

Ignorance is not bliss

We are witnessing a catastrophic loss of species that is the direct result of human activities. Yet we remain scandalously ill informed about the processes that give rise to biodiversity, and the consequences of its loss.

If variety is the spice of life, we face an increasingly bland future. There are perhaps 10 million species of organism on Earth, of which at most 1.8 million have been described. In some taxonomic groups, up to 20% of known species face extinction, and countless more are disappearing unnoticed. This should concern us all because we don't know what the consequences will be. In general, the less diverse an ecosystem, the less productive and stable it is. But ecologists are currently unable to make specific predictions that could help inform decisions about development and conservation.

If this is to change, we must reinvigorate taxonomy and describe the vast ranks of unnamed species. We need more passionate field workers, like Peter Ng of the National University of Singapore, whose efforts to catalogue neglected faunas are profiled on page 396. And we must ensure that the results of their endeavours don't languish on dusty shelves.

We also need to answer practical questions about the consequences of biodiversity loss. How many species are needed for an ecosystem to function? Will the loss of certain key species have disproportionate knock-on effects? This research must be done on appropriate scales of time and space: consider biodiversity over too short a time, or too small an area, and you can get the wrong answers.

Many interested scientists say gloomily that governments are not interested in this work. Given the stakes, this defeatism isn't good enough. Taxonomists and ecologists should look to the visionaries in their own midst, and to what their colleagues in genetics and climatology have achieved by understanding how to cast a research agenda in a light that can inspire — and if necessary, alarm — politicians.

Few have a clearer vision than Charles Godfray, director of the UK Natural Environment Research Council's Centre for Population Biology at Silwood Park, west of London. He argues that taxonomy must emerge from museums to become a web-based information science (H. C. J. Godfray *Nature* **417**, 17–19; 2002). Some initiatives of this ilk are under way, but the call has been short-sightedly rejected by much

of the taxonomic community, notably the Linnean Society of London.

Godfray was also instrumental in setting up one of the few long-term ecological projects investigating the consequences of declining biodiversity in a developing country where the problem is particularly acute. With backing from Britain's Royal Society, the Sabah Biodiversity Project in Malaysian Borneo is investigating ecosystem function and timber production in felled forests planted with varying numbers of species of dipterocarp — the main type of tree found in the rainforests of southeast Asia.

More projects of this type are needed, but they won't be forthcoming unless ecologists can take a leaf from the book of the geneticists whose lobbying in the late 1980s led to the Human Genome Project. There are parallels between the two research agendas. Like taxonomy, genome sequencing is purely descriptive, while the Sabah study of ecosystem function is conceptually related to systems biology, the probing of the function of gene networks that has followed in genomics' wake. Taxonomists and ecologists need to dispel the notion that their work — which involves dirty boots, rather than gleaming lab machinery — is somehow less scientific.

The cheerleaders of genomics promised gains in terms of human health and economic output. The economic consequences of ecosystem management are harder to quantify, but they are no less real: sustainable forestry, agriculture and tourism can all put developing economies on a sounder footing, to the benefit of us all.

Climatologists faced similar problems in explaining the economics of their case. After global warming was identified as a threat, some leading climatologists became highly effective lobbyists, pounding the corridors of power to stress the importance of their work. They won increased research funding and the establishment of the influential Intergovernmental Panel on Climate Change.

So far, taxonomists and ecologists have failed to muster a comparable response to the galloping loss of our planet's biodiversity. It's time that they did. ■

Drug research abused

Political pressures threaten to undermine a key agency involved in tackling the problems posed by drugs.

If you listen to the US government, drugs are bad — end of story. That shallow drag on a passing joint will launch you on a slippery slope towards social exclusion, jail and general misery. But if you listen to scientists instead, your outlook on drugs may be more nuanced. Not everyone becomes addicted, and exactly how drugs such as marijuana and ecstasy affect the brain over time, for example, are questions that only painstaking research can resolve.

In a rather uncomfortable spot between these two standpoints sits the US National Institute on Drug Abuse (NIDA) — tasked by Congress with providing the science that guides drug treatment and prevention, and under heavy pressure from its political paymasters to reinforce their war on drugs. Many critics say the agency's research agenda has occasionally been biased as a result (see page 394).

It takes a strong leader to maintain the integrity of research in the

face of such political pressure. Nora Volkow, the agency's director, radiates passion about science — and asserts that experiments, not politics, will guide her decisions.

A report last week from the Union of Concerned Scientists discusses the case of a researcher who, when being considered for a position on NIDA's advisory council, was questioned by a White House staffer about whether she supported "faith-based" drug treatment programmes, or voted for President Bush. This and other examples illustrate how the influence of politics has made itself felt in NIDA, and the tough task that Volkow faces to keep the agency clean.

NIDA's goal is to figure out the risks posed by drugs and how best to help those who suffer from taking them. It is not to reinforce politicians' ideas that we should lock up those who try drugs. When it comes to politics, NIDA should just say no. ■





Hazard warning
Biologists suggest formulae for listing dangerous microbes
p388



Drug dilemma
South African minister sparks row over AIDS treatment
p389



On the range
Native Americans angry over control of archaeological site
p391



Head start
Architect sets up dome on first-class stamp
p392

Los Alamos grinds to a halt after classified information vanishes

Rex Dalton, San Diego

In an unprecedented move, all research at Los Alamos National Laboratory (LANL) in New Mexico was shut down on 16 July. The drastic step follows the loss of computer-storage devices, which went missing earlier this month. Instead of doing research, many employees will this week be undergoing extra training while top government officials scrutinize the lab.

It has been a bad month for US nuclear-weapons labs — LANL is not the only one to have had security problems. On 15 July, Sandia National Laboratories in New Mexico, which conducts military research, reported losing a classified computer disk, although just a day later officials said that they had found it. But some observers and researchers say that LANL seems to be getting rougher treatment for its security lapses, and wonder if this might be an attempt to swing a future bid for lab management to a different institution.

The loss of the storage devices is LANL's third security failure in eight months. Last December an inventory revealed ten storage devices were missing, then in May it was found that a computer device had been destroyed without proper documentation. In the current case, two missing devices were found misplaced in a high-security area, but two others remained unaccounted for as of 19 July — all came from the weapons physics directorate.

Officials at the University of California (UC), which manages LANL, attribute the security failures to a "cowboy culture" in which some researchers fail to follow policies for managing classified material.

Top officials from the Department of Energy, which owns the labs, arrived this week to help find the missing devices and to rectify problems. They were accompanied by members of the House committee that oversees the national labs, including its chairman Joe Barton (Republican, Texas).

There are about 12,000 employees at Los Alamos, nearly half of whom are scientific staff. During the stand-down, LANL researchers will take inventories of classified material and undergo repeat training exercises. Some research may resume this week.



A series of security lapses at Los Alamos National Laboratory have put its management under scrutiny.

Meanwhile, UC officials say that they are intensifying efforts to instill a new culture in which security lapses won't be tolerated. Peter Nanos, LANL's director, wrote a blunt e-mail to staff to emphasize the point: "If you think the rules are silly, if you think compliance is a joke, please resign now." Although similar statements were made after the earlier gaffs, there is a fresh urgency to the comments this month.

Trouble at the top

The problems come at a bad time for UC, which has managed both LANL and the Lawrence Livermore National Laboratory — another nuclear-weapons lab — for decades. Earlier this year, the energy department decided to open the management of both labs to outside bids; proposals for managing LANL will be invited after November's presidential election. The role is coveted by others, including the University of Texas and defence contractor Lockheed Martin.

Congressmen from Texas — particularly Barton — have heavily criticized UC's management of LANL. At a hearing of the House Committee on Energy and Commerce last week, Barton said: "There's probably better security at the public library over CDs and videos." He also demanded that hundreds of lab employees take lie-detector tests.

But when the missing computer disk was announced at Sandia, which is run by Lockheed Martin, Barton was quiet. A spokesman for Barton told *Nature* that he had no comment on Sandia's problem at this time. Sandia's security lapse was spotted during an annual audit at the end of June, but wasn't made public for two weeks. Lockheed Martin officials say they are tightening procedures.

Such scenarios have prompted some UC professors to wonder whether congressional Republicans and Bush administration officials are trying to play up embarrassing management problems at UC-managed labs, perhaps to aid a Texas bid. Winning the contract to manage a major nuclear-research facility would be a great financial coup for any university. Over the past week, memos detailing past security issues at LANL have been leaked and issues raised about safety incidents, some of which seem routine. "The timing of these events makes people wonder," one LANL researcher told *Nature*.

UC will be trying hard to bring its staff in line with security rules over the coming months as researchers get back to work. Regardless of who controls the lab in the future, experts say that they will be keen to keep classified material where it belongs — not misplaced, missing or wreaking havoc with the lab's public image. ■

Biologists seek to revamp biowarfare register

Helen Pearson, New York

A rebellion is brewing among US scientists who handle pathogens that could be used in biological warfare. The official list of hazardous bacteria, viruses and toxins is so frustrating to some microbiologists that they are trying to rationalize it.

The 'select agents' list, drafted by the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, details some 40 microbes and toxins that could threaten public health if they escaped or fell into the wrong hands. Labs that work with them must follow strict security measures. But some researchers are unhappy with the register, pointing out that relatively innocuous pathogens, such as the fungus *Coccidioides immitis*, are placed under the same tight restrictions as lethal viruses such as smallpox. "The list is very arbitrary," says microbiologist Arturo Casadevall of the Albert Einstein College of Medicine in New York.

Critics say that the list was drafted in haste after the 2001 anthrax mail attacks, without sufficient consultation with the scientific community. They say it relies too heavily on whether agents have been used as biological weapons in the past, rather than their potential for use in the future.

Now Casadevall and his colleague Liiseanne Pirofski have suggested a way to redo the list by calculating the biological-weapon potential of each microbe (*Trends Microbiol.* 12, 259–263; 2004). Their formula considers a pathogen's virulence, ability to jump from person to person, incubation period and capacity to remain in the environment. The



Fears of a biological attack have led to strict research controls.

CDC's list takes similar factors into account, but not in a quantitative fashion.

The new system could be used to partition microbes into categories reflecting different levels of threat. At the moment, "you're either on the list or you're off it," says Reynolds Salerno of Sandia National Laboratories in Albuquerque, New Mexico. It could also be used to work out whether emerging infectious diseases should be added. The virus responsible for severe acute respiratory syndrome (SARS), for example, which is not on the list, comes out with a weapon potential above that of anthrax and below that of smallpox.

"The formula is a good attempt to bring order to the process," says Ian Lipkin, an expert

on infectious diseases and biodefence at Columbia University in New York.

As yet, there is no agreement on the precise formula that should be used. Salerno is working on his own formula, which factors in how easily a dangerous strain might be obtained by bioterrorists and the simplicity with which it could be converted into a bioweapon.

Many researchers have mixed feelings about seeing their favourite organism labelled a select agent. On the one hand, they fret about strict security requirements that can hamper their research. On the other, such a listing can help bring in cash, as a separate but overlapping list is used by the National Institute of Allergy and Infectious Diseases (NIAID) in Bethesda, Maryland, to allocate biodefence-research funds.

The existence of such similar, overlapping lists further frustrates researchers. The lists are often confused with each other, says Casadevall — he admits that even the published version of his paper managed to mix up two of the lists without the reviewers or the journal noticing.

The CDC should recruit a panel of experts from the scientific community to resolve these issues, says Janet Shoemaker, head of public affairs at the American Society for Microbiology, who has worked closely with the government over the regulation of biological agents.

For the moment, the CDC says that it has no plans to revise its select-agent list. The NIAID says it plans to talk with other government agencies about possibly using a formula to bring at least some of the various lists into line with each other. ■

Budget delays threaten to leave US science in limbo

Geoff Brumfiel

While researchers relax for the summer holidays, science lobbyists in Washington are feeling tense about the state of the US budget.

For the third straight year, the budget looks as though it will be severely delayed — raising concern that research may be stymied in the months to come.

As *Nature* went to press, Congress had failed to complete work on any of the 13 spending bills that provide funding for the US government in fiscal year 2005, which begins on 1 October 2004. Squabbling over the overall budget has held up work on the bills.

With the Congress summer recess set to begin on 23 July, this situation seems unlikely to change much. Experts say little will probably be done until after the

November presidential election. Many think that any shift in power will make it difficult to resolve the budget bills before the newly elected members of government take over in January. "I can't imagine we will be done until February or March," says one staffer at the House of Representatives.

That could mean trouble for science funding, according to Janet Shoemaker, head of public affairs at the American Society for Microbiology in Washington. If the budgets aren't passed on time, Congress will pass a series of 'continuing resolutions' that will supply funding at 2004 levels. Until a budget is passed, agencies will be unable to fund new projects set to begin in 2005, Shoemaker says. Although most agencies anticipate a few months' delay each year, the

potential for a prolonged hold-up is worrying. "Beyond January it becomes a severe problem," she says.

That was the case after the mid-term election in 2002, when the government ran on continuing resolutions for nearly five months before finally approving a budget in February (see *Nature* 421, 774; 2003).

Work done on the budget so far indicates that science agencies are unlikely to get much in the way of new funds even when the spending bills do pass. A report last week from the House subcommittee working on the budget for the National Institutes of Health indicated that the agency could receive an increase of \$727 million for 2005 — a 2.6% rise from its current funding level of \$28.5 billion. ■



Bitter pill: an antiretroviral that protects babies can leave their mothers resistant to treatment.

Health minister ignites row over drugs for HIV mothers

Erika Check, Bangkok

Doctors should provide more treatment for pregnant women with HIV — not less. So argued scientists last week, after a controversial statement by the South African health minister at the XV International AIDS Conference in Bangkok.

The row focused on studies of a treatment that reduces the risk of women transmitting HIV to their babies during childbirth; some 700,000 children were born with HIV last year. Scientists already know that the treatment — a single dose of a drug called nevirapine (sold as Viramune) given to the mother during labour — can encourage drug-resistant viruses to grow in the mother's body. On 11 July, Gonzague Jourdain, with his Thai and French colleagues, announced that women with such drug-resistant viruses respond less well to the antiretroviral nevirapine when it is used to treat their disease later in life. The team published its work last week (Jourdain *et al.* *New Engl. J. Med.* **351**, 229–240; 2004).

South African health minister Manto Tshabalala-Msimang told delegates that her country would stop recommending nevirapine based on this evidence. Scientists called her statement misguided. A sudden ban of the drug would unnecessarily cause disease in children, they say, because no other treatment is available in most poor countries. They add that the week-long controversy has overshadowed the real importance of the work, which could save the lives of mothers and children.

"This is not the Titanic sinking, where we have to save the mother or the child," says Marc Lallemand, an author of the study from the IRD, a Paris-based research institute for sustainable development. "We are putting lifeboats all around, so everyone — mothers and children — can get out."

The women who were most affected in Jourdain's study were also the least healthy to begin with. So the lesson, says Jourdain, is that sickly mothers should be given better care before labour, which could include taking a drug in combination with nevirapine that reduces both the chance of the baby getting HIV and the chance of resistance developing in the mother. "We should improve what is in place already," says Jourdain, who works at the Harvard School of Public Health.

Another team of researchers at the conference presented similar conclusions about resistance in South African women. Infectious-disease specialist James McIntyre of the University of Witwatersrand in Johannesburg reported that if women were given nevirapine before labour and a two-drug combination therapy just afterwards, far fewer carried strains of virus resistant to nevirapine — resistance was cut at least fivefold. Jourdain called the study "preliminary, but encouraging".

In Bangkok, world health officials spoke out against Tshabalala-Msimang's stance on nevirapine. By the end of the week, the South African government had decided that although it did not recommend using nevirapine alone, it would not immediately stop dispensing it. On 14 July, the World Health Organization said that single-dose treatments should not be undermined, but alternatives should be considered where possible.

All of this spells good news for mothers and their unborn children. But many at the conference still worry about the possible effects of Tshabalala-Msimang's comments on drug policies in Africa and Asia. "The confusion our minister sows is undermining prevention programmes around the world," says activist Zackie Achmat of the South African group Treatment Action Campaign. ■

Stem-cell specialists split over proposal for US repository

Jonathan Knight, San Francisco

A plan to set up a national embryonic stem-cell bank in the United States has met with a decidedly mixed response. Some are calling it a political move that does nothing to address what they see as real problems with the Bush administration's policy on stem-cell research. But others hope it will prepare the National Institutes of Health (NIH) for possible future changes in policy.

The proposal, announced by Health and Human Services Secretary Tommy Thompson on 14 July, would bring all the embryonic stem-cell lines currently available to US researchers under one roof. This includes 19 current cell lines, along with a handful that are not yet ready for distribution.

Embryonic stem cells can theoretically be coaxed to generate any tissue in the body, and could one day provide treatments for diseases such as diabetes and Parkinson's. The bank would grow the cell lines in standardized conditions, aiming for uniform quality, lower costs and easier access for researchers.

"I think this is just the right kind of organization and service that the NIH should provide to the community," says Fred Gage, a stem-cell researcher at the Salk Institute for Biological Studies in La Jolla, California.

But critics point out that the bank does nothing to address the fact that there are currently so few stem-cell lines available for study. The Coalition for the Advancement of Medical Research (CAMR), a Washington-based pressure group, says the move is a worthless attempt by the Bush administration to appear active in stem-cell research. "It's a bit like giving a new paint job to a car without an engine. It's still not going to take you where you need to go," said CAMR spokesman Sean Tipton in a statement released last week.

Strangely, current distributors of stem cells do not seem to have been consulted before the announcement. Meri Firpo, who runs the distribution centre at the University of California, San Francisco, only heard the news when contacted by *Nature*.

And Andy Cohn, a spokesman for the WiCell Research Institute in Madison, Wisconsin, which distributes 5 of the 19 approved lines, said he did not yet know how the bank would affect his organization's work. "We don't know what the plan is," he says. ■

Space group touts 'next steps' for astronauts

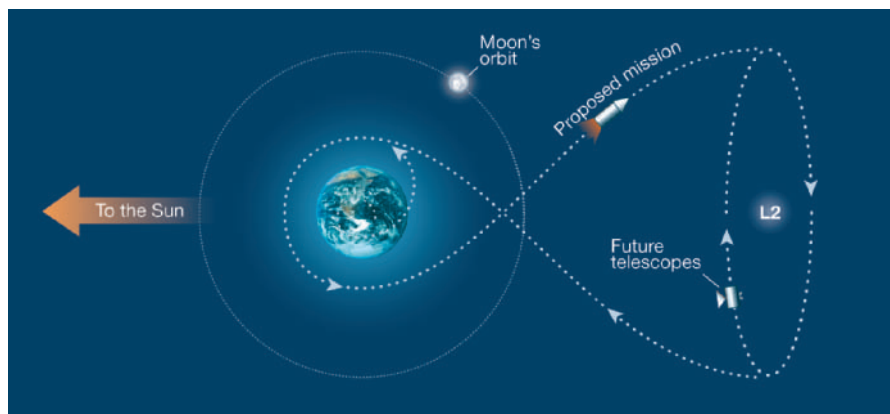
Tony Reichhardt, Washington

Think more broadly and don't forget international cooperation: that is the message from a veteran group of space scientists and engineers. Together they presented an alternative to NASA's vision for human space exploration at a Paris press conference on 19 July.

Their report, called *The Next Steps in Exploring Deep Space*, caps a three-year, informal study sponsored by the International Academy of Astronautics. Although it was a volunteer effort with no official government backing, the study reflects the judgement of more than 100 space professionals. These include former NASA space-science chief Wesley Huntress and Bernard Foing, project scientist for SMART-1, Europe's first lunar mission, which was launched last year.

NASA is making plans for a return to the Moon by 2020, to act as a stepping stone to Mars. But the academy suggests two other short-term destinations for astronauts: a Sun–Earth libration point (L2), where future astronomical telescopes will be stationed (see *Nature* 419, 666; 2002), and near-Earth asteroids. The group recommends that the decision about which goal to tackle first should be based on scientific objectives, and that each successive step should require only one major technology to be developed, partly to lower costs.

The authors call their study "an example of what could be done, not a prescription of what will be done". Yet they clearly intend to broad-



Over the Moon: some say the Sun–Earth libration point L2 is a good target for human missions.

en the 'where next?' debate beyond what some feel is too narrow a focus on the Moon.

The construction and maintenance of astronomical facilities at L2 (see above, not to scale) may provide a good rationale for a programme of human exploration beyond low-Earth orbit, the study's authors write. Astronauts could also be sent to explore asteroids — collecting data to aid in protecting the Earth from possible future collisions.

Huntress, now director of the Carnegie Institution of Washington's Geophysical Laboratory and the study's leader, says its recommendations are in line with President George W. Bush's directive to send humans beyond Earth orbit. But the committee adds

that its alternative priorities for space exploration make more economic sense. Making L2 — a point in space four times farther away than the Moon — an early goal for astronauts would defer the cost of a lunar lander and base until after a deep-space vehicle is proven, for example.

The academy places stronger emphasis on the possibility of international cooperation than did a recent US presidential commission. So far, international discussion of Bush's Moon–Mars programme has produced only polite interest. The European Space Agency is conducting its own study of possibilities for human space exploration, but this is not expected to be finalized until next year. ■

Britain decides 'open access' is still an open issue

Declan Butler

Can journals function if authors, instead of readers, carry the cost of publication? An inquiry by the UK House of Commons Science and Technology Committee concluded this week that we will just have to wait and see. After five months of investigating access to journals in science, technology and medicine, the committee has reported that the concept of 'author-pays' open access seems "viable" but requires "further experimentation".

In the meantime, the report advises the government to oblige UK authors to publish articles on their institutions' websites.

Many people have questioned whether the author-pays open-access model, as pursued by the Public Library of Science and BioMed Central, for example, is economically sustainable. At the same time, the current system of 'reader-pays' has resulted in spiralling journal costs that many libraries can no longer afford. "This cannot continue," says committee chairman Ian Gibson, a Labour member of parliament.

The report gives advice on running open-access schemes more smoothly. For example, it suggests that funders include money in grants to cover author fees.

But the report also says that it is too early to tell how open access will pan out. "The author-pays model needs more work; that's why we are saying we shouldn't go into it right away. We have to look at the possibilities and perhaps have a pilot scheme for a certain length of time," says Gibson.

Most of the data used in the debate, such as the cost of publishing, come from the 'grey literature' of reports from the UK-based charity the Wellcome Trust, and from publisher statements or online debates, says Gibson. "I'm suspicious of the figures thrown around," he adds. The committee recommends that the government carries out a comprehensive independent study.

Its strongest recommendation is that the UK government should ensure that funders make it compulsory for researchers to post their papers online. "Our idea — a rabbit out of the hat — will make the university library

system sit up and listen," says Gibson.

The idea of posting material online has been around for a decade, and an increasing number of institutions are building online repositories. DSpace, for example, developed at the Massachusetts Institute of Technology, aims to store the institute's entire intellectual output, including data and course materials (see *Nature* 420, 17–18; 2002).

Gibson says he hopes the report will make researchers aware of the issue. "The sad thing is that academics don't really care as long as they get their work published," notes Gibson. According to a recent survey by the Centre for Information Behaviour and the Evaluation of Research at City University London, 82% of working scientists say they know little or nothing about open access. ■

▶ nature.com/nature/focus/accessdebate

For more news and analysis go to
news @ nature.com
www.nature.com/news



Rich pickings: this ranch in Range Creek features a wealth of undisturbed archaeology.

Native Americans voice fears for relics in land-transfer deal

Rex Dalton, San Diego

Archaeologists are celebrating a 'find of a lifetime' at a Utah ranch packed with ancient relics. But some Native Americans in the area are upset that they were not consulted about how the land should be treated when it first passed into public hands for study some three years ago.

The 1,700-hectare ranch, in Range Creek, Utah, sports countless archaeological sites, most of which are undisturbed. The previous owner of the property, Waldo Wilcox, was well aware of the scientific value of the land — he and his family had maintained the site for more than 50 years. A shotgun-wielding Wilcox would warn off intruders and looters venturing onto his property.

But while rumours of the archaeological value of the site swirled in academic circles for years, it wasn't until 2002 that Utah state archaeologist Kevin Jones toured the ranch and saw its true riches — including dwellings, granaries and rock art dating from between 2,000 and 700 years ago.

Others have likewise been impressed. "I was blown away," says Duncan Metcalfe, curator of archaeology at the Utah Museum of Natural History in Salt Lake City and leader of the site survey. "I've never seen anything like this in my 25 years of archaeology," he adds.

Researchers are now planning studies and consultations with Native American tribal leaders about any future excavations. Such consultations are designed in part to accommodate their beliefs on the handling of human remains, which are often at odds with archaeological study. In the case of the Kennewick Man skeleton found in Washington state, for example, tribes have legally delayed research for eight years for cultural reasons.

But some Native Americans in the Range Creek area — including Melvin Brewster, a Northern Paiute with a doctorate in archaeology — are angry that these consultations are coming so late. They point out that the land has been in government hands for years. In 1998, an organization of hunting and fishing enthusiasts petitioned for federal legislation to buy the ranch, ensuring its availability for sport. The US Bureau of Land Management (BLM) then took title to the land for \$2.5 million in 2001, conveying it to the state of Utah in January 2004.

Brewster claims that the Utah deal violated federal laws on historic-site planning, environmental policy and protection of Native American remains. Calling the land conveyance an "under-the-table deal", Brewster claims that the BLM had an obligation during the years of federal ownership to consult with tribes about Native American remains and to seek public comment on any plans for the ranch. "The American public has been violated," he says.

Brewster says he worries that Native American relics may not be treated properly now that the land is owned by the state of Utah. He is concerned that insufficient funds may be available to stop looting, for example.

Don Banks, BLM spokesman for the Utah region, says that the deal was done based on special legislation secured by two congressmen from Utah. He adds that no one sought any special legal advice about the transfer.

For now, both Native Americans and scientists are excited about the discoveries expected from the ranch. "I have faith that Utah will lay out a plan to work with all of us," says Leigh Kuwanwisiwma, director of the Hopi Cultural Preservation Office in Arizona. ■

Project probes impact of waste carbon dioxide on marine life

David Cyranoski, Tokyo

Japanese researchers are beginning to make public the first data from a project that could allow waste carbon dioxide to be dumped in the ocean. The tests are aimed at finding what concentrations of carbon dioxide can be tolerated by worms and bacteria living in the muddy sea floor.

But critics of the scheme say that even the 'safe' limits determined by the project could still be harmful in the long term.

Researchers and industry are interested in pumping waste CO₂ down to the sea floor to reduce atmospheric levels of the greenhouse gas. But the idea is controversial. Recent tests have shown that the rise in oceanic CO₂ levels from natural absorption from the air over the next century could have a detrimental effect on sea creatures' ability to make shells (R. A. Feely *et al. Science* 305, 362–366; 2004). Higher concentrations from dumped CO₂ could have an even greater impact.

In the project, conducted by the Research Institute of Innovative Technology for the Earth (RITE) in Kyoto and the research arm of the Kansai Electric Power Company in Osaka, plastic containers were placed on the sea floor off the coast of Japan. Carbon dioxide was then pumped into the containers.

Extremely high concentrations of 20,000 and 5,000 parts per million caused most visible creatures, such as nematodes, to die, the team says, although the abundance of smaller creatures and bacteria was barely affected. Shigeo Murai, the ocean chemist who heads the project, says that RITE is working on technologies to dilute CO₂ to concentrations tens of thousands of times lower than those used in the experiments before introducing it to the sea floor. This means it should have no effect on marine life, he says.

But Kenji Kato, a microbial ecologist at Shizuoka University, says that low concentrations could also have an effect in the long term. Even if the number of bacteria stays the same, he says, the species of bacteria present could be changing: "An entirely new community could grow in a few months."

It is also unclear how the CO₂ would disperse once it is on the sea floor. Experiments to assess this have been blocked in the past by environmental protesters (see *Nature* 419, 6; 2002). Murai plans to look at a natural outlet of CO₂ seeping into the ocean floor around Okinawa to study its dispersion. ■

Public puts faith in nanotech despite little knowledge

Washington Many people on both sides of the Atlantic have little idea what 'nanotechnology' means — but they still think it will benefit them in the long run.

A poll of more than 1,500 US adults, backed by the National Science Foundation, has found that 80% of people have heard little or nothing about nanotechnology. Nevertheless, 40% of respondents felt that it would bring more benefits than risks, and a further 38% felt the risks and benefits would be about equal.

A British survey released in March by the Royal Society and the Royal Academy of Engineering found similar levels of ignorance: 71% of UK respondents had never heard of nanotech. But of those who knew the term, 68% thought it would make things better in the future.

The surveys are likely to reassure scientists working in the field, who have become alarmed at activists' attempts to portray their work as a threat to human health and the environment (see *Nature* 424, 246–248; 2003). But the US poll has a sting in its tail: 60% of respondents said they do not trust business leaders to minimize the technology's risks.

Europe joins forces to study cell-division genes

London The genes and proteins that orchestrate the division of human cells are to be investigated in an €8.5-million (\$10.6-million) project funded by the European Commission. The MitoCheck project will involve scientists at the European Molecular Biology Laboratory in Heidelberg, Germany, and ten other research institutes.

The researchers plan to sift through some 20,000 human genes using RNA interference to silence them one by one, and observe the regulatory effects that individual genes have on the cell cycle. "To do this for a single gene is a routine procedure, but to do this for tens of thousands of genes with precision is a very challenging task," says project coordinator Jan-Michael Peters of the Research Institute of Molecular Pathology in Vienna, Austria.

Buckminster Fuller gets stamp of approval

Washington The polymath Richard Buckminster Fuller is honoured this month with a US first-class stamp. The image, taken from a 1964 cover of *Time* magazine, depicts Buckminster Fuller's head as a geodesic dome, his best-known creation.

After inventing gadgets such as a three-wheeled car and modular bathroom, Buckminster Fuller patented the geodesic dome in 1954 as a means of making strong, lightweight buildings. His domes are still not widely used in architecture, but their structure was found in the 1980s to be echoed in a series of ball-shaped carbon molecules, now known as the fullerenes.

♦ www.bfi.org



Cell biologists hope that the study will also reveal the mechanics of how proteins are modified by phosphorylation as cells undergo division.

♦ www.mitocheck.org

Arizona braces itself for dengue invasion

San Diego As the rainy mosquito season looms, there are growing fears that dengue fever may be about to break into the southwest United States. An explosion of cases in Mexico, near the border with Arizona, has researchers scrambling to find out why the mosquito-borne virus is spreading north.

In 2003, there were 1,177 cases of dengue fever in the Mexican state of Sonora, compared with just 258 in 2002. Cases of the most serious form of the disease — dengue haemorrhagic fever, which kills in 5% of cases — climbed from 98 in 2002 to 346 in 2003.

Researchers at the Center for Insect Science at the University of Arizona in Tucson, working with Mexican scientists, have found that this upsurge of cases has coincided with the emergence of the DEN-4 strain of the virus, typically seen farther south, in addition to the DEN-2 and DEN-3 strains. This is a concern, because the haemorrhagic form of the disease occurs when people are serially infected with more than one viral strain.

The last instance of dengue infection in mainland United States was in 1999, when 18 people were found with the disease in Laredo, Texas.

Graduate students cannot form unions, board rules

Washington Graduate students at private US universities have been stripped of their right to unionize. The National Labor Relations Board ruled on 13 July that graduate students at Brown University in Providence, Rhode Island, should not be allowed to join a union as they are students and not employees.

The decision reverses a ruling by the board in October 2000 that allowed graduate students who had research or teaching duties at New York University to unionize (see *Nature* 408, 123; 2000). That decision was made while the board had a majority of Democrats. President George W. Bush has since appointed three Republicans.

The ruling bars collective bargaining at private schools such as Brown, but will not affect state-run public universities. Some states, including California and New York, do allow graduate students to form unions.

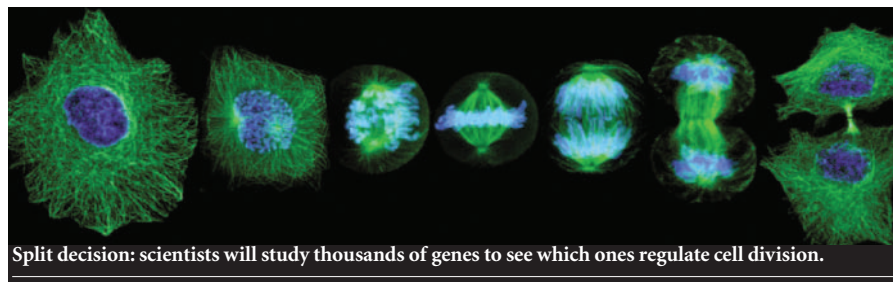
Feeding frenzy paves way for crazy ant death

Sydney Work has begun to save Australia's Northern Territory from one of the world's most voracious pests. Billions of yellow crazy ants will be poisoned in an attempt to halt the insects' trail of devastation.

Yellow crazy ants (*Anoplolepis gracilipes*), so-called because of their chaotic movements, originated in Africa and are one of the world's most invasive species. Instead of forming queens that fly away from the parent colony to establish new communities, they form dense supercolonies with up to 1,000 ants per square kilometre of bush.

In Australia's northeast Arnhem Land, the insects have now infested 25,000 square kilometres of land — feasting on local flora, and killing or out-competing resident invertebrate populations.

Scientists plan to combat the ants by dropping granules of poison onto the colonies from helicopters. The pellets, called Presto, contain a fishmeal product that the ants love but other animals avoid.



Split decision: scientists will study thousands of genes to see which ones regulate cell division.

A hard habit to break

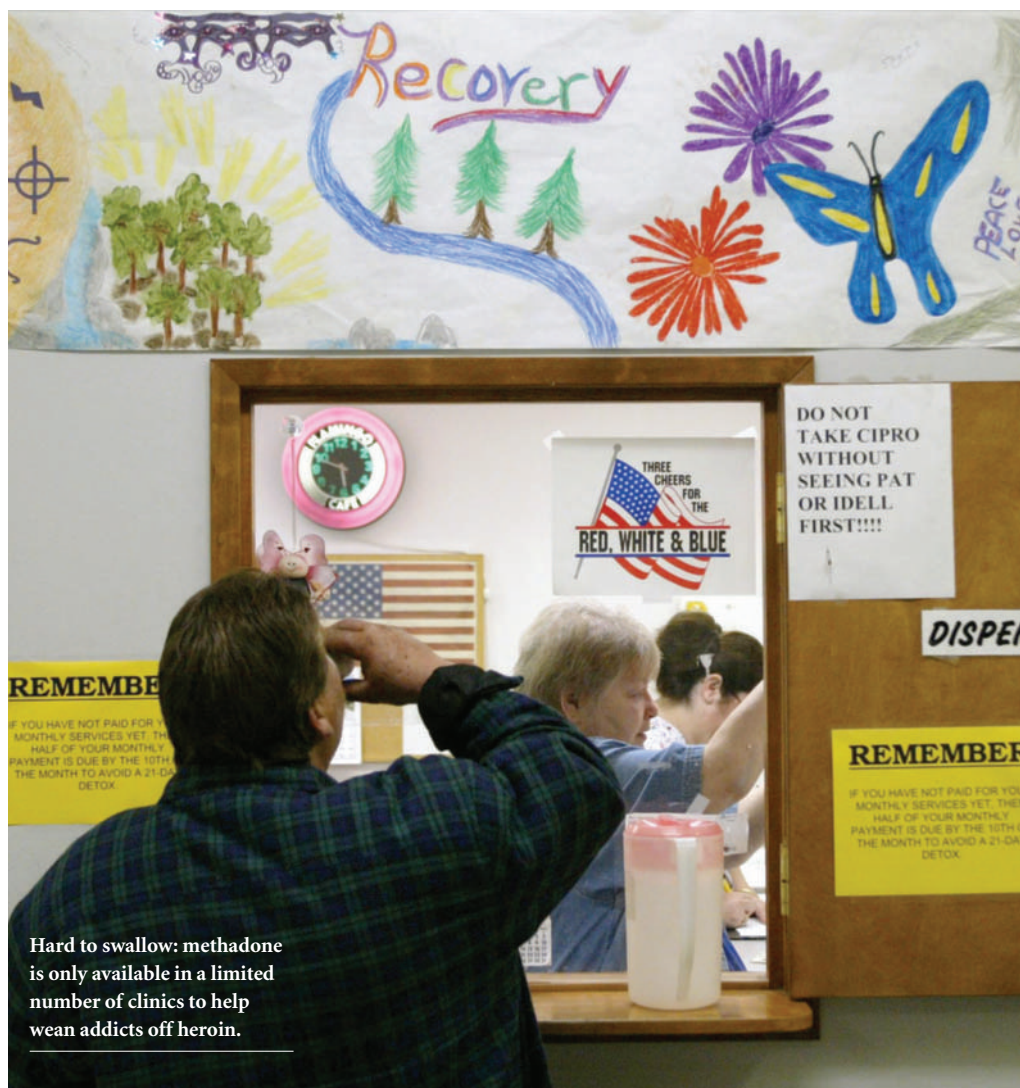
The US National Institute on Drug Abuse has frequently been accused of bowing to the political agenda of its paymasters. But, as Helen Pearson finds out, its new director swears that the agency is being led by science.

One chilly Thursday morning in December 1999, Alan Leshner, then head of the US National Institute on Drug Abuse (NIDA), shared a stage at a Washington DC press conference with a series of striking, multicoloured brain scans.

At the top of his poster, a set of yellow and orange images showed a brain packed with healthy neurons that communicate using the chemical serotonin. Beneath them, a matching set of scans from the brain of a long-time ecstasy user revealed dark, gaping holes — purportedly illustrating the havoc wreaked by the drug.

If would-be ecstasy users found the images alarming, so did some scientists. Earlier that year, a few researchers had alleged that there were serious flaws in the 1998 study¹, led by George Ricaurte at Johns Hopkins University School of Medicine in Baltimore, Maryland, from which the images were taken. They felt that problems with the positron emission tomography experiments that generated the images may have exaggerated the loss of serotonin neurons in the brains of ecstasy users².

Critics of NIDA say that the use of the images typifies the agency's sometimes-cavalier approach to research. They charge that its outlook is overly influenced by the 'war on drugs' launched by President Richard Nixon in 1971, and pursued relentlessly by US politicians ever since. "NIDA's agenda has been profoundly shaped by a drug-war ethos," says Craig Reinerman, a sociologist who studies drug policy at the University of California, Santa Cruz.



Hard to swallow: methadone is only available in a limited number of clinics to help wean addicts off heroin.

But the agency's present director, Nora Volkow, rejects this characterization, arguing that the agency's mission is driven by scientific impartiality. "I'm a scientist, not a politician," she says, "and my value is to be able to provide objective information."

Opening salvos

Like the war on drugs, NIDA was born amid growing concern about drug abuse during the hippy era. Founded by Congress in 1974, it was initially part of a now-defunct branch of the US health department, which dealt with alcohol, drug abuse and mental health. In 1981, Congress gave individual states control over treatment and prevention of drug abuse and NIDA became a research agency; only later, in 1992, did it become part of the National Institutes of Health (NIH).

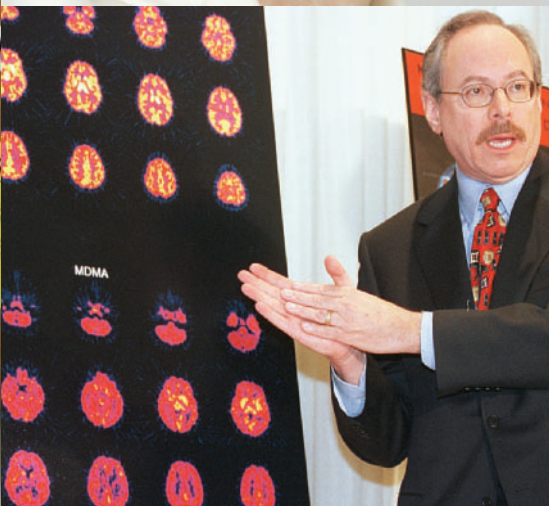
The agency now has an annual research budget of nearly \$1 billion, the bulk of which is distributed in grants to researchers at universities and medical schools. Most of them credit NIDA with funding high-quality research.

But accusations that politics usurps science at the agency have never been far

beneath the surface, and they peaked under Leshner's 1994–2001 tenure. "Leshner was a minister of propaganda in the war on drugs," says John Morgan, a pharmacologist who studies medical marijuana and drug policy at the City University of New York. Leshner, now executive director of the American Association for the Advancement of Science in Washington, declined to be interviewed for this article.

NIDA's critics level three main charges against the agency. First, they contend, it tends to support research projects that will document the terrible damage caused by drugs — and so bolster the government's view that these substances are unsafe. "It's science in the service of politics," claims Charles Grob, who studies hallucinogens at the University of California, Los Angeles.

Second, they argue that NIDA neglects the investigation of medicinal uses of recreational drugs, or of anything else that might show the substances in a better light. "I got the message that it will fund things showing harm, but when it comes to benefits there's no chance," says Alan Marlatt, a psychologist at the University of Washington, Seattle, who



Nora Volkow (top) may answer critics who claim that her predecessor, Alan Leshner (above), paid too much attention to the war on drugs.

served on NIDA's National Advisory Council on Drug Abuse from 1995 to 2002.

Third, they allege that NIDA shuns research into 'harm reduction' — approaches that seek to minimize the death, disease and social damage caused by drug use, rather than eliminating usage outright. At NIDA "it's just say no — or nothing at all", says Marsha Rosenbaum, director of the San Francisco office of the Drug Policy Alliance, which campaigns for changes in national drug policy.

But in May 2003, the Bush administration did something that might yet assuage the agency's harshest critics. It appointed Volkow, a neuroscientist then working at Brookhaven National Laboratory in New York state, as director. The choice has been widely applauded. "With Nora Volkow at the head I'm far more optimistic about the direction NIDA is going in," says Grob.

Volkow is a wiry, intense woman who runs six miles a day and munches her way through chocolate to keep her sugar levels

up. Her forthright manner and outstanding work in brain imaging had already marked her out as a rising star in US science. In a 75-minute interview in her office in Bethesda, Maryland, she argued emphatically that her agenda is being driven by science.

Volkow says that she was brought up with a natural wariness of politics: as the great-granddaughter of the Russian revolutionary Leon Trotsky, many members of her family were assassinated — including Trotsky himself, in Mexico in 1940. "I guess it made me reluctant to participate in the political process," she says.

And she brushes off the suggestion that NIDA is under political pressure to support particular kinds of research projects. "I have not had a scientist come to me and say: 'I'm afraid if I submit it, it'll be rejected,'" she says. During her tenure, she adds, "there has never been an instance where a grant got a good review and then was not funded because of political incorrectness". The only political constraint that she has to deal with, she notes, is the one faced by all NIH institute heads — she has to convince Congress of the importance of NIDA's work in order to obtain funding.

Compromised position

The critics, of course, see this as part of the problem. They think that the agency is already steeped in a culture sympathetic to the war on drugs and that Volkow too, will eventually succumb. "I think she's getting the message that if you want to keep your money, you do what they say on the hot-button issues," says Ethan Nadelman, director of the Drug Policy Alliance in New York.

As for the suggestion that NIDA's approach to drug research is too narrow, Volkow points out that research into the medicinal uses of marijuana, for example, clearly falls outside the agency's mission, which is "to bring the power of science to bear on drug abuse and addiction".

It is the third area of criticism — that NIDA habitually neglects a wide swathe of approaches to drug abuse that are being explored outside the United States — that Volkow struggles to dispel. Critics say that the agency has failed to put sufficient resources into investigating, for example, the effectiveness of methadone as a legal substitute for heroin. One problem, they say, is that NIDA has not examined whether patients would benefit if methadone, which is currently available only at designated clinics, were made available through general doctors' offices and pharmacies. "The range of research questions being asked is remarkably narrow," says Nadelman.

NIDA has also steered clear of trials that would prescribe heroin itself in the course of

weaning hardcore addicts off the drug. Such trials have produced tentatively promising results in other countries: in 2001, a study by Jürgen Rehm of the Addiction Research Institute in Zurich, Switzerland, for example, tracked nearly 240 patients who had been prescribed the drug. Of those in the programme for at least 18 months, the number with severe mental-health problems halved and many stopped stealing and began therapy aimed at kicking the drug completely³.

On the basis of this and other studies, a consortium of scientists called the North American Opiate Medication Initiative (NAOMI) drew up plans in 2001 to start a multi-site clinical trial of prescription heroin in Canada and the United States. But would-be US participants say that they realized it was a non-starter after speaking to NIDA officials. They think that this was because it was politically unacceptable for the agency to fund research that delivered illegal drugs to addicts. "It was a foregone conclusion," says Ernie Drucker of the Montefiore Medical Center, New York, who was part of the NAOMI consortium.

Several researchers interviewed by *Nature* back Drucker's view. They say that there is an unspoken rule that some types of proposal are not worth submitting to NIDA. To gain funding, they say that they frame proposals in ways that emphasize the damaging effects of drugs, or that leave out contentious phrases such as 'harm reduction'. "There are certain words you don't put in the title of your grants," says William Miller, who studies addictive behaviours at the University of New Mexico in Albuquerque.

Volkow says that she is not intrinsically opposed to research into harm reduction, but would place other priorities, such as investigation of a newer alternative to methadone, called buprenorphine, above the prescription of heroin.

She vows to maintain the separation between science and politics at NIDA. Since she took over, she says, groups that she declines to identify have pressed her to issue a public statement that marijuana causes brain damage, on the basis of her own imaging studies. Volkow says that she has refused, because it is not yet established that the patterns shown by these studies actually affect health or behaviour. "I don't want to use science to scare," she says, "I want to use science to educate."

Helen Pearson works in New York for *Nature's* online news team.

1. McCann, U. D., Szabo, Z., Scheffel, U., Dannals, R. F. & Ricaurte, G. A. *Lancet* 352, 1433–1437 (1998).
2. *Lancet* 353, 1268–1271 (1999).
3. Rehm, J. et al. *Lancet* 358, 1417–1420 (2001).

♦ www.nida.nih.gov

M. CAVANAUGH

M. SMITH/NEWSMAKERS



A tragedy with many players

Peter Ng is a man with a mission: to catalogue the huge diversity of life dwelling in habitats long dismissed as uninteresting. It's a race against time, he tells Carina Dennis and Peter Aldhous.

A tropical peat swamp is not a welcoming place. Its acidic waters sting every tiny scratch on your body. Hold your hands just beneath the surface and you can't see them through the tannin-laden water. The only bonus is that leeches don't fancy the murk. But Peter Ng, a taxonomist and conservation biologist at the National University of Singapore, loves getting up to his armpits in the mire.

Ng has discovered that the peat swamps of southeast Asia are teeming with rare species of fish and crustaceans, many of which are new to science. "Peat swamps have been badly neglected," says Ng, who pulls out novel specimens on nearly every dip into these hostile waters. His team has found a treasure trove of biodiversity in other unlikely places too, including the broken rubble of dead coral found off tropical beaches.

Now Ng is engaged in a race to catalogue these neglected faunas before many of them are wiped out by Asia's relentless economic development. The peat swamps, in particular, are being drained as fast as he can sample them, sometimes for urban or agricultural development, at other times — in a bitter irony — under the guise of 'environmental improvement'.

The rich faunas found in such neglected habitats underline a growing realization that conservation biologists really know very little about our planet's biodiversity (see 'Hyperdiversity, or hype?', overleaf). But Ng's quest is driven by more than academic interest. "Scientists and environmental managers need to know these habitats exist and that they deserve to be conserved," he says. "If they are lost through ignorance or misinfor-

mation, then it will be a terrible tragedy."

Ng's excitement about peat swamps was first fired in the early 1990s, when his surveys of the North Selangor swamp in Malaysia turned up some unexpected finds. Wading through the muck, his team stumbled on several species of an elusive genus of catfish called *Encheloclarias*¹. The last time anyone reported finding such a fish was in the 1930s.



Swamp king: Peter Ng has discovered dozens of species new to science.

The pH of such swamp waters can be as low as 3 — about the same as vinegar. "Until recently, peat swamps were assumed to be hostile, acidic places where the biodiversity was low," says Ng. "But that's because no one had actually jumped in." After taking the plunge into numerous swamps on the Malay Peninsula and the islands of Borneo and Sumatra, Ng's team has found some 80 fish new to science, bringing the

estimate of the total number of species in the swamps to 200–300. "My students say they have too much to study," he says. A high proportion of the species are exclusive to the peat-swamp environment².

Hide and seek

It isn't easy pulling these creatures from the mire. Many of the fish and crustaceans hide in crevices in the peat, rather than swimming in the watery channels that run through it. Ng has to wade into the murky water holding a net, while his larger colleague, Maurice Kottelat, jumps up and down on the banks. "Maurice does his rain dance," says Ng — and that's enough to startle the creatures out of the peat. If Ng wants to sample all of the species present, including amphibians, he has to collect at night as well. Wading in a swamp in pitch darkness isn't for the easily spooked, he says.

"I'm interested in fish in places where people haven't been," says Kottelat, who is president of the European Ichthyological Society and is based in Cornol, Switzerland. Some of the fish they have found in the swamps — including colourful species of *Betta* fighting fish, popular with the aquarium trade³ — have thrown up some intriguing mysteries. "Why they should be so colourful in such black water is the million-dollar question," Ng says.

Ng doesn't spend all his time in the field wading chest-deep in fetid swamps. Sometimes his trips take him to much more pleasant destinations, including the idyllic sandy beaches of the US-controlled island of Guam, some 2,000 kilometres east of the Philippines. There he studies another overlooked habitat: the piles of broken coral rubble that lie just a few tens of metres from the shore.

If you swim from a tropical beach to its fringing reef, you first pass over expanses of this rubble, created by storms and the ravages of time. Many naturalists make this journey time and again, lured by the stunning array of fish and other animals in the reef itself. But few pause to give the rubble a second thought. "It looks like a desert, almost," says Ng.

The rubble piles might still be dismissed





What lies beneath? Acidic, fetid and unfriendly, southeast Asia's peat swamps (above) are home to a rich diversity of life, including unique and surprisingly colourful species of *Betta* fish (opposite).

as a desert, were it not for the obsessive curiosity of Harry Conley, who began diving off the beaches of Guam in the early 1980s after retiring from the US Air Force. Conley's initial interest was in collecting seashells. Realizing that they were present in large quantities in the rubble, he started to excavate the piles by hand. "The holes he dug looked like bomb craters," says Ng.

Rubble rousers

By the 1990s, Conley's interests had expanded to the crabs and other creatures that his digs disturbed. He teamed up with researchers at the University of Guam led by Gustav Paulay, who was conducting a survey of the island's marine biodiversity⁴. Ng joined the team in 2000, when he was asked to help identify crabs pulled from the rubble.

Sadly, Conley's work came to an abrupt end in 2002 when he was killed by a gunshot to the head — probably self-inflicted — after an argument in which another man was shot and wounded. Friends describe Conley as a generous yet troubled individual. "Harry was

always a loner," recalls Bruce Henke, who accompanied him on many collecting trips. "He was very competitive. If you found something, he had to find something better."

Henke, who works in Guam as an air-traffic controller, continues to collect samples from the rubble piles. Now 50, he has dived recreationally since his teens and is an accomplished underwater photographer. In shallow waters, each dive might last for two hours, consuming several tanks of air. Larger pieces of rubble have to be prised away with a crowbar. "It can be dangerous," says Henke. "It's not something that I would recommend to just anybody."

Like the peat swamps, the rubble piles fascinate Ng because many of the species present have been found in no other environment. The crabs alone include about half-a-dozen genera and dozens of species that were previously unknown to science⁵. Those that live deep within the rubble look similar to crabs that are found in caves,

"These places don't have big, sexy animals. But in almost all cases, when they say a place is species-poor, they're wrong."

— Peter Ng

with no pigments and shrunken eyes⁶. "It's a whole new world," says Ng.

Paulay, who is now at the University of Florida in Gainesville, believes that the rubble piles support a similar range of creatures to those that live in crevices deep within the coral reef itself. "This 'cryptofauna' represents the bulk of the reef's biodiversity," argues Paulay, who views the rubble as a window through which to study creatures that otherwise could be reached only by destroying the reef.

Hidden depths

There could be even more life inside the rubble than has been found so far — including faster-moving creatures that dart away from the dig sites before they can be caught. "I would expect to find prawns, gobies and worms of all sorts," says Ng. Sampling the sites more thoroughly is a huge challenge, though. One way would be to deploy a heavier-duty version of the vacuum pumps that marine archaeologists use to suck up silt to sift through for fragments of artefacts. But such devices don't yet exist.

Other parts of the marine ecosystem, such as the steep walls that fall away from the reef into deeper waters, are even harder to reach. Scuba divers can't descend below 100 metres and the walls are too steep to be dredged. Without access to a submersible, the only means of sampling the area is to throw a net blindly into the depths and cross your fingers.

Guam's rubble fauna isn't under immediate threat, so there is opportunity to think about how to improve sampling. But time is running out for many of Ng's field sites, which include extensive systems of limestone hills and caves found across southeast Asia. Water seeping down from the dense forest there creates moist grooves and crevices in the rock that host a wide assortment of plants and animals.

These 'karst' systems haven't been entirely overlooked — they include the Gunung Mulu National Park in Malaysian Borneo, granted World Heritage status for its biodiversity and geological features. But



Pincer movement: these crabs are among the many species that make their home in piles of smashed dead coral.

their faunas are poorly studied⁷. “Limestone hills are hotspots of richness, but this has only really been appreciated recently,” says Tony Whitten, a biodiversity specialist with the World Bank, based in Washington DC.

The forests that grow above the limestone are rapidly being cleared, which dries out the rock, destroying its subterranean ecosystem. Ecologists studying the caves often find themselves side-by-side — literally — with developers or cement companies mining the hills for raw material. “We’ll be carefully collecting our samples, trying not to damage the habitat, and a few hundred metres further on, there’s a guy busy with a chainsaw,” says Jaap Vermeulen, a collaborator of Ng’s at the Singapore Botanic Gardens, who specializes in the taxonomy of snails.

Vanishing world

Living in the concrete jungle of Singapore, Ng knows all too well the consequences of unfettered development. Last year, he published a paper in *Nature*⁸ that used the detailed records made by British colonial naturalists to document the extinctions that have occurred since most of Singapore’s forests were cut down. Extrapolating from these data, Ng and his colleagues concluded that up to 42% of the species currently in southeast Asia’s forests will disappear over the next century if habitat destruction continues at its present rate. About half of these will be global extinctions, as the species are not found elsewhere.

The race to catalogue biodiversity before it disappears is particularly intense in the peat swamps, which are disappearing at a frightening rate. The drainage is even affecting neighbouring bits of forest; dried peat bogs have fuelled huge fires that have razed some areas.

When he talks about the threats to the peat ecosystems, Ng’s natural enthusiasm can’t hide a deep melancholy. He has become



Coral rubble used to be ignored, but it offers researchers access to creatures that live deep within a reef.

a reluctant ambulance chaser, rushing in to sample sites earmarked for development. The collecting methods that Ng’s team uses in such cases are severe and destructive. “We call them salvage operations,” he says. “We catch whatever is scientifically valuable, knowing full well that there is no tomorrow. It is a very rotten feeling.”

Ng hopes his work will counter the ignorance that underlies the blasé destruction of these habitats. Officials and developers argue that there is no point conserving the swamps because there is ‘nothing’ there. “These places don’t have big, sexy animals,” Ng concedes. “But in almost all cases, when they say a place is species-poor, they’re wrong.”

He remains gloomy about the chances of protecting the remaining swamps from the tide of development. But at the very least, Ng is determined to reveal for future generations the true magnitude of the devastation

that is now being wrought. “The story is much more tragic once you know the characters,” he says.

Carina Dennis is *Nature’s* Australasian correspondent; Peter Aldhous is *Nature’s* chief news & features editor.

1. Ng, P. K. L. & Lim, K. K. P. *Ichthyol. Explor. Freshwaters* 4, 21–37 (1993).
2. Ng, P. K. L. *Wallacea* 73, 1–5 (1994).
3. Kottelat, M. & Ng, P. K. L. *Ichthyol. Explor. Freshwaters* 5, 65–78 (1994).
4. Paulay, G. *Micronesica* 35–36, 3–25 (2003).
5. Paulay, G., Kropp, R., Ng, P. K. L. & Eldredge, L. G. *Micronesica* 35–36, 458–517 (2003).
6. Ng, P. K. L. & Takeda, M. *Micronesica* 35–36, 419–432 (2003).
7. Vermeulen, J. & Whitten, T. *Biodiversity and Cultural Property in the Management of Limestone Resources — Lessons From East Asia* (World Bank, Washington DC, 1999).
8. Brook, B. W., Sodhi, N. S. & Ng, P. K. L. *Nature* 424, 420–423 (2003).
9. Boucher, G. & Lamshead, P. J. D. *Conserv. Biol.* 9, 1594–1604 (1995).
10. Lamshead, P. J. D. et al. *BMC Ecol.* 3, 1 (2003).
11. Lamshead, P. J. D. & Boucher, G. *J. Biogeogr.* 30, 475–485 (2003).
12. Gray, J. S. *Mar. Ecol. Prog. Ser.* 112, 205–209 (1994).

Hyperdiversity, or hype?

Back in the mid-1990s, John Lamshead was a cheerleader for a radical idea: that the mud of the ocean’s depths is every bit as rich, in terms of biodiversity, as a typical rainforest. Today, he has a different message.

Lamshead, an expert on the taxonomy of nematodes at the Natural History Museum in London, has found that the conclusions you draw about biodiversity can vary dramatically, depending on the area over which you measure it.

In 1995, he and Guy Boucher of the Paris Natural History Museum described the diversity of nematodes — worm-like creatures that account for about 80% of the abundance of animals in seafloor mud — recorded in nearly 200 sediment cores from around the world. They counted up the number of species found in each batch and extrapolated from these samples to estimate that

the deep sea is a ‘hyperdiverse’ environment, containing more than a million species⁹.

But this extrapolation was controversial. The cores had been analysed by different people, so it was unclear how much overlap there was between the species found in different cores. To tackle this problem, Lamshead’s PhD student Caroline Brown analysed cores taken along a 3,000-kilometre stretch in the central Pacific¹⁰. Her results showed that, although every core was very diverse, a similar pattern of species repeated itself from sample to sample. In retrospect, says Lamshead, this shouldn’t be that surprising: explorations by research submersibles have revealed that the deep-sea environment varies little from place to place.

Based on Brown’s evidence, Lamshead and Boucher officially downgraded their estimate of

the diversity of the deep ocean in a guest editorial in the *Journal of Biogeography*¹¹ last year. Lamshead is now doing further work with John Gray of the University of Oslo in Norway, who had independently questioned the idea of deep-sea hyperdiversity¹². Their unpublished data confirm that the same picture applies to polychaete worms — the next most abundant creatures in the mud.

Lamshead’s repudiation of deep-sea hyperdiversity has met with some resistance from marine ecologists, who had eagerly embraced the concept. Environmentalists had also used the theory to argue that it is crucially important to conserve every segment of the deep ocean. Conservation is still important, says Lamshead, and the deep ocean still diverse — just not hyperdiverse.

P.A.

The animal-care regulatory system is a sham

Inspectors often overrate an experiment's value and underestimate the pain it causes.

Sir—Your News story “Britain seeks compromise on animal research” (*Nature* **428**, 882; 2004) focuses on the UK government's proposals to increase investment in non-animal research as a means of addressing public concern with the scientific and ethical validity of animal experimentation.

Although such a policy may have some long-term effect in reducing the level of suffering experienced by animals and producing better research methods, there are other steps that could be taken immediately to address problems with the regulatory framework surrounding animal research.

Your passing reference to the overwhelming lack of trust in the current regulatory regime hits the nail on the head. The number of government inspectors is tiny in comparison with the scale of

research, and the majority of those inspectors have backgrounds in animal experimentation.

Work by our campaigning group, Uncaged, has shown that, instead of acting as neutral arbiters, inspectors often share the viewpoint of animal researchers. When granting licences, they seem to underestimate animal suffering, and/or overestimate the usefulness of the research.

Pig-to-primate xenotransplantation research conducted by Imutran at the Huntingdon Life Sciences laboratories in Cambridge, between 1994 and 2000, provides a particularly stark example. Leaked confidential documents published online by Uncaged in April 2003 (see www.xenodiaries.org/evidence.htm), following legal battles with Imutran, reveal that procedures leading to the collapse and death of higher primates were classified by

Home Office inspectors as of merely ‘moderate’ severity — despite the government's stated policy that these procedures should be classified as ‘substantial’ or ‘severe’.

Our legal battle to publish the inside story of this research was won on the basis of our claim that the documents reveal the Home Office's failure to enforce its own rules. Yet there are no constitutional mechanisms for ensuring accountability, and the government continues to ignore a call by 153 members of parliament for an independent inquiry, as reported in *The Guardian* newspaper on 11 November 2003.

The public is absolutely correct to distrust our sham of a ‘regulatory’ system.

Dan Lyons

Uncaged Campaigns, 9 Bailey Lane, Sheffield S1 4EG, UK

Label of ‘autism’ could hold back gifted children

Sir—As a person with autism, I would like to respond to the ongoing discussion, by Allan Snyder (*Nature* **428**, 470–471; 2004) and Oliver Sacks (*Nature* **429**, 241; 2004), about the links between autism and genius.

In my book *Thinking in Pictures* (Doubleday, New York, 1995) I was one of the first people to suggest that Einstein had traits of an adult with mild autism. He had no speech until the age of three. The statement I mistakenly attributed to Oliver Sacks was from Swedish psychiatrist Christopher Gillberg, who wondered whether or not the composer Béla Bartók or the philosopher Ludwig Wittgenstein may have been autistic (*An Anthropologist on Mars*, Knopf, New York, 1995, p 295). Sacks described Gillberg as “one of the finest clinical observers of autism”, and he seemed to agree with the above statement.

I agree with Oliver Sacks that the terms autism and Asperger's syndrome are overused. I give talks at many autism conferences. When the milder diagnosis of Asperger's syndrome became popular in the 1990s I started seeing many intellectually gifted children at these conferences. I told one mother that, before Asperger's syndrome became widely accepted, her child would have received a label of ‘intellectually gifted’.

In my work I have observed many engineers and scientists who seem to have mild autism or Asperger's traits. One of my

great concerns is that a child diagnosed with Asperger's or mild autism will be held back by the label and not have their talents (scientific or otherwise) developed.

Temple Grandin

Department of Animal Science, Colorado State University, Fort Collins, Colorado 80523, USA

Keeping a clear head on effects of illicit drugs

Sir—It is surprisingly hard to find a level-headed, fact-based discussion of issues related to illicit drugs, as is unfortunately illustrated by your recent Editorial “Think harder about ecstasy” and News Feature “The ups and downs of ecstasy” about MDMA or ‘ecstasy’ (*Nature* **429**, 113 & 126–128; 2004).

According to your Editorial, “MDMA can cause psychosis, hyperthermia and even death in some people who take the drug recreationally. But there is no research to indicate whether or not this will be a problem in the controlled settings of a clinical trial”. This last statement ignores many published phase I studies, including some mentioned in the News Feature.

At least five independent research teams around the world have administered MDMA to more than 200 human subjects in controlled experimental settings, and neither hyperthermia nor psychosis — let alone death — has ever been a problem (see ref. 17 from your News Feature, and others available on request). Increases in body temperature in clinical settings have

never exceeded 1 °C, with some studies failing to find any significant increases. Likewise, no cases of psychosis have been reported in controlled clinical trials with pre-screened subjects. Although there may be low levels of individual symptoms (such as thought disturbances or perceptual illusions reported by Vollenweider, myself and colleagues in 1998; see ref. 13 of the News Feature) that can also occur in clinical psychoses, these disappear completely after drug effects have subsided, and they are a far cry from full-blown psychotic states.

It therefore seems misleading for your Editorial to mention only the 12 documented cases of psychosis related to recreational ecstasy use, and for the News Feature to list only “moderate thought disorder” when citing findings from our 1998 study concerning the psychological effects of MDMA. In fact, this study, like others, found that MDMA produced several effects, including positive mood, perceptual alterations and slight anxiety over loss of control that was not psychotic.

Non-selective reporting of factual information is the basis for responsible decisions on drug issues, and is in the interests of both sides of the debate on the therapeutic use of MDMA.

Alex Gamma

University Hospital of Psychiatry, Lenggstrasse 31, 8029 Zurich, Switzerland

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter.

A question of balance

How safe are the medicines that are prescribed to children?

Paul Ramchandani

All parents worry about their child getting sick. When the illness is depression, it is difficult for parents to know what to do for the best. Doctors and parents of depressed adolescents have been alarmed over the past year by the news that certain antidepressant drugs (selective serotonin reuptake inhibitors, or SSRIs) are no longer recommended for use in under-18s. The concerns arose following reports of unpublished clinical data related to these drugs, and last month the pharmaceutical company GlaxoSmithKline was sued for allegedly suppressing negative findings from clinical trials for its SSRI drug paroxetine (also known as Paxil or Seroxat). These events raise serious questions about the reporting of research and the transparency of the drug regulation system. Many commentators are now asking for the results of all clinical trials to be made available for public scrutiny, and for trials conducted on children this seems an inescapable conclusion. Unfortunately, at present, we cannot know how many other child medications are affected by the selective disclosure of results.

How did we let this happen? The bodies that license medicines in many countries require pharmaceutical companies to disclose the results of all relevant clinical trials if they are applying for a licence to market a drug for a particular disease. But children and adolescents are not always protected by this safeguard, as many drugs used to treat children are used 'off-licence' — that is, they are not licensed for a particular condition in children, but can be used if a doctor recommends it. Estimates from the United Kingdom suggest that at least 50% of all drugs used to treat children in hospitals are off-licence¹. Except for child-specific conditions, trials of medications in children are less common, as they are often difficult to conduct, and so less is known about how effective or safe many medicines are for children.

For conditions that affect both adults and children, such as diabetes, asthma, depression and anxiety, drugs that are originally licensed for adults gradually begin to be used in young people. This 'trickle-down' prescribing can occur in many countries which do not require a separate licence for prescription to children. This means that differences between the effects of the medication on adults and children may not be detected, and can also prevent independent scrutiny of clinical trial data.

At present, we cannot be sure how widespread this problem is. Given the lack of



Out in the cold: concerns over drug safety could make treating children for depression more difficult.

transparency in the system, other drug trials may have been conducted in children and not reported. This problem is slowly being recognized. In 2002, the United States introduced a 'paediatric rule', whereby pharmaceutical companies can get an extended drug licence if they collect more information about the drug's safety in children and release it to the US Food and Drug Administration (FDA). It seems that GlaxoSmithKline was applying for paediatric use of paroxetine when the unpublished data came to light. How this information was handled by the regulators remains unclear. Of course, the best way of ensuring independent scrutiny is to make the data publicly available.

Therapies for the young

Among the many voices in this debate, the ones we hear from least are the young people facing depression, their families and the doctors who treat them. Depressive disorder is estimated to affect 2–6% of adolescents, causing low mood and energy, and loss of interest in usual activities. Suicidal thoughts and attempts at suicide are not uncommon. The two most important treatments for this age group are a form of talking therapy called cognitive behaviour therapy (CBT), and antidepressant drugs — predominantly the SSRIs. Published research^{2,3} had previously

suggested that these drugs were effective, well tolerated treatments for depression in adolescence, and they have been used increasingly in recent years.

Throughout 2003, re-evaluation of the evidence underpinning the use of SSRIs in depressed children and adolescents by the UK Committee on Safety of Medicines (CSM) led to growing restrictions on their use in young people. This affects a very large number of children and adolescents who take these drugs (40,000 in the United Kingdom and many more worldwide). The CSM review was prompted partly by concerns that these drugs might increase the risk of suicidal thinking in children and adolescents. At present, only one SSRI, fluoxetine (better known as Prozac), is left without contra-indication, for use on adolescents in the United Kingdom — although it is not actually licensed for that purpose. In the United States, the FDA has also restricted the use of paroxetine, and is currently reviewing the data relating to all other SSRIs. It is expected to report later this year.

The CSM analysis revealed two startling facts. First, several important clinical trials had never been published in medical journals, or previously scrutinized by the CSM, despite being completed some time before. The unpublished trials were usually negative in outcome (that is, they found either that the

TRIGHT/CORBIS

SSRI drug was no more effective than a placebo, or that there were higher levels of side effects). To paraphrase a recent analysis⁴ in *The Lancet*, the good news about the drugs was released into the public domain, but the bad news was not. Even last month, a paper was published⁵ that reported a clinical trial for the SSRI drug citalopram, suggesting that it improves depressive symptoms. It was surprising to see that the authors of that paper were either unaware or chose not to refer to a larger unpublished trial of the same drug, which suggested that there was little evidence for efficacy in 13–18 year olds, even though details of this other trial had been released in the CSM report last year⁶.

Second, problems have been observed with the way in which results have been interpreted in the published clinical trials: a recent analysis published in the *British Medical Journal* suggested that in many instances the benefits may have been exaggerated while the adverse effects of the drugs have been played down⁷. On the whole, this has tended to bias the way in which these drugs have been evaluated, leading to a more positive view than was justified by all of the available information. This raises uncomfortable questions. The opacity of the current regulatory system has not helped. In the case of paroxetine, the CSM claimed to have received new data only last May — years after some of the trials were completed. The FDA has not publicly stated when it received the data, although the manufacturer, GlaxoSmithKline, claims that it complied with all existing requirements to distribute the results. We still cannot be sure who had what data, when; and what steps were taken to analyse them.

These shortcomings in the reporting of SSRI data are serious. But equally important is the ongoing treatment of depressed children and adolescents. Many newspaper headlines interpreted the CSM's warnings as an outright ban, and implied that the drugs are ineffective or dangerous for depressed adolescents. But the picture is more complicated than that. The CSM has since clarified its recommendations, explaining that the drugs should only be used when supervised by specialists (child and adolescent psychiatrists) under more restrictive conditions. Nonetheless, such confusion may make it more difficult to prescribe SSRIs,

even when they could potentially help a severely depressed adolescent.

Most medical treatments are a trade-off between the potential benefits and harms, and these are different in different people. Three points are worth making here. First, there are no reported suicides in the data analysed by the CSM. There is also only limited evidence linking SSRIs with an increased risk of suicide attempts or of suicidal thinking, although there are other side effects. Three different analyses of the data^{4,6,8} have so far reached differing conclusions (see Table), and a fourth, commissioned by the FDA, is yet to report.

In addition, many clinicians are concerned that the clinical trials have largely been conducted on groups including much younger children and adolescents, whereas



Commonly prescribed antidepressants.

— in the United Kingdom at least — older adolescents with very severe depression would be those most likely to receive medication. There have not yet been any trials to see whether older adolescents respond differently to the drugs than younger children, which might be expected as the antidepressants seem to work in adults. Additionally, in several trials, those with suicidal thoughts or believed to be at high risk of suicide (and so likely to be the most depressed) are excluded^{2,3}. So the findings from trials may be unrepresentative of those most likely to receive the medications in clinical practice.

Third, although the CSM took, perhaps unavoidably, a black-and-white view, restricting use of most SSRIs, but leaving fluoxetine as 'okay', the differences between the drugs are more marginal than that would suggest. Given the relatively small differences in effectiveness between fluoxetine and some of the other antidepressants, and the limitations

of the existing clinical data, most child psychiatrists would still consider using some of the other SSRIs. For example, when treating older adolescents who are severely depressed, or who do not initially respond to CBT, an SSRI may be worth trying.

More generally, how do we ensure that the results of all clinical trials become publicly available? To an outsider, it seems there are too few checks and balances. Analysis of clinical trials approved by ethical review boards show that less than a third of completed research is published within three years⁹, despite an ethical obligation to do so. It has long been acknowledged that there are few incentives for authors or journals to publish negative, or even inconclusive, findings. In response to the SSRI controversy, a group of top medical journals is considering asking drug companies to register all clinical trials in a public database at inception, as a prerequisite for later publication. Other proposals would require companies to reveal data from all completed trials on marketed products, whether they study licensed or unlicensed uses. Although companies may have resisted such openness in the past to protect themselves commercially, this argument should not be permitted when the health of individuals is at stake.

A few pharmaceutical companies have said that they will move in this direction, having signed up to good-practice guidelines (www.gpp-guidelines.org), and an independent resource for registering trials already exists (www.controlled-trials.com). GlaxoSmithKline has also announced that it will provide summaries of all clinical trials in an online database, although it is unclear whether this will include making all trial data available for independent scrutiny. Although these individual moves are welcome, the industry as a whole needs to adopt new standards of transparency. The regulatory bodies, the ethics committees that approve these trials, the doctors involved in trials, and the journals that publish them, also have a role to play. Together, we must ensure that when doctors and patients make a decision to use a drug, they do so knowing they have access to all the information they need. ■

Paul Ramchandani is in the Department of Psychiatry, University of Oxford, Oxford OX3 7JX, UK.

Risk of suicidal thinking and behaviour in children taking SSRIs for depression

Drug	CSM advice (ref. 6)	Whittington <i>et al.</i> (ref. 4)	Brent and Birmaher (ref. 8)
Paroxetine	Evidence of increased risk	Inconclusive evidence	No significant evidence*
Fluoxetine	No significant evidence	No significant evidence	No significant evidence
Sertraline	Evidence of increased risk	Inconclusive evidence	No significant evidence
Citalopram	Evidence of increased risk	Inconclusive evidence	No significant evidence
Venlafaxine	Evidence of increased risk	Evidence of increased risk	No significant evidence

*No statistically significant evidence at $p < 0.05$ level.

- Aynsley-Green, A. *et al.* *Br. Med. J.* **321**, 229–232 (2000).
- Keller, M. B. *et al.* *J. Am. Acad. Child Adolesc. Psychiatry* **40**, 762–772 (2001).
- Wagner, K. D. *et al.* *J. Am. Med. Assoc.* **290**, 1033–1041 (2003).
- Whittington, C. J. *et al.* *Lancet* **363**, 1341–1345 (2004).
- Wagner, K. D. *et al.* *Am. J. Psychiatry* **161**, 1079–1083 (2004).
- http://medicines.mhra.gov.uk/ourwork/monitorsafequalmed/safetymessages/ssrioverview_101203.htm (accessed 28/06/04).
- Jureidini, J. N. *et al.* *Br. Med. J.* **328**, 879–883 (2004).
- Brent, D. & Birmaher, B. *J. Am. Acad. Child Adolesc. Psychiatry* **43**, 379–380 (2004).
- Pich, J. *et al.* *Lancet* **361**, 1015–1016 (2003).

Acknowledgements

Iain Chalmers, Rosie Nicol-Harper, Alan Stein and Stuart Logan provided helpful comments.

Journey of a scientific hero

A fictionalized biologist finally gets the biography he deserves.

Beyond the Outer Shores: The Untold Odyssey of Ed Ricketts, the Pioneering Ecologist Who Inspired John Steinbeck and Joseph Campbell

by Eric Enno Tamm

Four Walls Eight Windows; 2004. 384 pp.

\$26

Jon Christensen

Ed Ricketts is a hero among marine biologists — mention his name and their faces light up. His 1939 field guide, *Between Pacific Tides*, is still in print and is often one of the first books a fledgling marine biologist takes into the field. Outside marine biology, however, few people recognize his name. But Doc from John Steinbeck's novel *Cannery Row*, that's another story. Almost everyone knows Doc.

"It is nearly impossible to separate the man from the myth, Ed Ricketts from Doc," writes Eric Enno Tamm in *Beyond the Outer Shores*. But Tamm does rescue Ricketts from the myth and gives us an account of Ricketts' heroic journey as a scientist who had his finger on the pulse of ecology through his intimate knowledge of the Pacific coast. Ricketts went to the University of Chicago, but did not graduate, before coming west to Monterey, California, to set up shop as a freelance scientist and purveyor of tidal-pool specimens. But he was a doctor of sorts. He diagnosed the environment and humanity, and offered humbling advice that his contemporaries failed to heed and that we are only now beginning to hear. He warned of the dire consequences of overfishing sardines, and realized that even his "slight bit of collecting" in the Great Tide Pool in Monterey probably contributed to the disappearance of some rare brittle stars.

Beyond the Outer Shores is the first biography of Ricketts. Tamm wisely focuses on the collecting trips that Ricketts took north and south along the Pacific coast, and on his friendships with Steinbeck and the great mythologist Joseph Campbell. The quest to understand the Pacific coast was central to Ricketts' life and work. By following that story, Tamm's book is likely to be the best biography of Ricketts for some time, even if others come along to provide more salacious tidbits from the party atmosphere that pervaded Ricketts' Pacific Biological Laboratories, both in Steinbeck's fiction and in real life.

Tamm has a personal interest in this story. He grew up in Canada on the remote western shore of Vancouver Island. In high-school biology class, Tamm surveyed the same beach



Novel research: Ed Ricketts was immortalized as the character Doc in John Steinbeck's *Cannery Row*.

in his village that, he would later discover, Ricketts had studied on his expeditions north to the outer shores. Tamm worked in the fishing industry and became a journalist and an environmental activist. He writes with an impassioned but sure hand. His book is meticulously researched and he has a firm grasp of the material.

It is easy to become lost in Ricketts' own obtuse scientific and philosophical writings, much of it never published in his lifetime. As Tamm puts it, Ricketts' "grammar and sentence structure were as tangled as a kelp bed". And Tamm could easily have gone overboard in siding with scholars who have overplayed Ricketts' profound influence on Steinbeck. But Tamm confidently sets a steady course through these shoals and brings readers along with him on the fascinating journey.

I know what it's like trying to follow Ricketts. This spring I set out from Monterey with a group of scientists from Stanford University's Hopkins Marine Station to retrace an expedition that Ricketts and Steinbeck took to the Gulf of California in 1940 to explore the intertidal zone; the duo collected more than 600 species, some 60 of which were previously undescribed. We went to see how things have changed.

In the book that resulted from their journey, *The Sea of Cortez*, Steinbeck and Ricketts noted that, in addition to sea stars, brittle stars and urchins, they had taken "2,160 individuals of two species of beer". Try as we might to match their enthusiasm, we only managed 1,200 individuals of six species of beer. But we found the gulf profoundly

changed. In "this region of the sea turtle and flying fish", we saw plenty of flying fish, but no turtles. We saw no great schools of tuna and big sharks, as our predecessors had. Instead, we found a new top predator, the jumbo Humboldt squid, *Dosidicus gigas*. And although the invertebrate diversity in the tide pools is still astonishing, it is lacking in species that people eat and collect, such as scallops and *Murex*, which one crew member of Steinbeck and Ricketts' expedition collected by the bushel for girlfriends back home.

"There is more of the whole man, John Steinbeck, in *The Sea of Cortez* than in any of his novels," wrote one critic in a review of the book that is still in print as *The Log from the Sea of Cortez* (although without the species catalogue that made up half the original book and shamefully with only Steinbeck's name on the cover now). "This is at once the record of a serious biological expedition and of the impact of a biologist and a novelist upon each other's minds... The best of Steinbeck is in it."

And the best of Ricketts, I would invariably have added before I read *Beyond the Outer Shores*. Now the best of Ricketts is in Tamm's biography. I read the book when I couldn't sleep for thinking of all the preparations still to be done before our departure to the Sea of Cortez. The thrilling sense of discovery on each page kept me up even longer. I read it again when we returned; it seemed even better the second time around.

Between Pacific Tides, Ricketts' guide to the "good, kind, sane little animals" of the California coast, remains his greatest individual

achievement — with help from friends and followers who cleaned up his prose and kept it updated through five editions. He never finished the grand survey that he dreamed of writing about the entire Pacific Coast, from Baja California to British Columbia. On 8 May 1948, a train hit Ricketts at a crossing near Cannery Row, as he was preparing for an expedition north to the outer shores with Steinbeck. He died three days later.

Ricketts' work was never published in scientific journals. His prescient analysis of the crash of the California sardine fishery was published in a local paper. His best ideas about ecology and humanity found voice in the writings of his friends, Steinbeck and Campbell. Now the whole life and work of Ed Ricketts can be found between the covers of *Beyond the Outer Shores*. ■

Jon Christensen is a Steinbeck fellow at San Jose State University, San Jose, California 95192-0090, USA.

Peas and helices

Mendel's Legacy: The Origin of Classical Genetics

by Elof A. Carlson

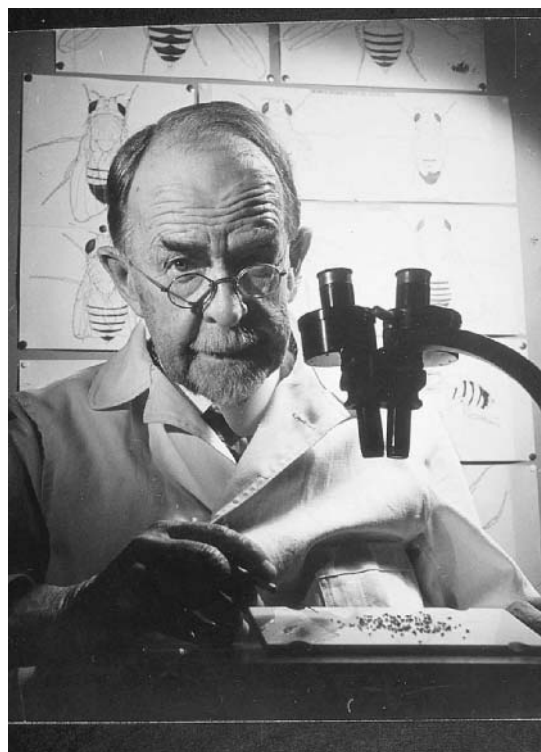
Cold Spring Harbor Laboratory Press: 2004. 352 pp. \$45

Garland E. Allen

Elof A. Carlson has added another important book to his list of publications on the history of genetics. *Mendel's Legacy*, in one sense a follow-up to Carlson's much earlier but still very useful *The Gene: A Critical History* (1966), represents the culmination of his work so far on the historical origins of classical genetics in the first half of the twentieth century.

Carlson defines classical genetics both chronologically and conceptually. He covers the first half of the twentieth century from the rediscovery of Mendel's work in 1900 to the publication of Watson and Crick's paper on the structure of DNA in 1953, and encompasses a number of strands of scientific thought going back into the nineteenth century. He justifies ending his story with the discovery of the double helix on the grounds that none of the preceding work based on the mendelian and chromosome theories led directly to discoveries about the molecular basis of heredity. Nor has classical genetics been replaced, in Carlson's view, by molecular genetics. Rather, the two have been integrated such that, although the molecular details of genetics are quite different from anything classical geneticists could have imagined, genes remain the fundamental elements of heredity, evolution and development.

In depicting the course of classical genetics, Carlson devotes chapters to the history



Model work: T. H. Morgan studied the genetics of fruitflies.

of evolution, cytology, embryonic development and agricultural breeding (principally the hybridization tradition in which Mendel worked). He asks a crucial historical question: if the roots of classical genetics lay in scientific traditions that were all developed in Europe, why did the United States become the place where mendelian genetics developed most rapidly and, in some ways, most successfully? Part of the answer clearly lies in the strong agrarian base in the United States at that time, with growing governmental support of agricultural research stations and the application of science to practical results.

Another part of the explanation, Carlson argues, is that the structure of the US university system was less encumbered with rigid, top-down professorial hierarchies than those in Europe. Once classical genetics had taken off, with T. H. Morgan and his group's work with the fruitfly, *Drosophila melanogaster*, and R. A. Emerson's group working on maize, the theory flourished by combining the cytological study of chromosome mechanics with breeding and hybridization experiments.

Carlson devotes a fair amount of space to cytology and cytogenetics, but includes the major players from all aspects of the work, including Hugo De Vries, Wilhelm Johannsen, Herman Nilsson-Ehle, W. E. Castle, E. B. Wilson, Barbara McClintock, J. B. S. Haldane and many others. *Mendel's Legacy* contains good thumbnail descriptions of the critical observations and experiments that went into the formation of the classical theory (more so, for example, than Robert Kohler's *Lords of the Fly*) while still incorporating

aspects of Kohler's study of social dynamics in the Morgan group. There is also a chapter on the fusion of mendelian genetics and darwinian evolution (the evolutionary synthesis), and glimpses into the problematic relationship between classical genetics and development in a discussion of E. B. Lewis's work on the homeotic gene complex (bithorax) in *Drosophila*.

Several outstanding features of this book will make it useful for specialists and non-specialists alike. One is the attempt to show how classical genetics was involved with political issues in the twentieth century, using three examples: eugenics (1883–1945), the Lysenko controversy in the Soviet Union (1930–1960), and the controversy over the genetic effects of radiation (1946–1970). Another noteworthy feature is the use of chronological tables for the whole field (at the beginning of the book), as well as for specific sub-topics, such as contributions

to the chromosome theory of sex determination, or geneticists' educational background. A third valuable feature is the author's liberal use of high-quality illustrations, including photographs of many geneticists seldom pictured before, original figures from published papers, and the title pages of important papers and books. The publishers, Cold Spring Harbor Press, have produced an attractive and useful book.

By its nature, *Mendel's Legacy* has to be selective, so issues of the philosophical basis of the gene, on which much has been written in recent years, are omitted. This is a pity, particularly because the issues that underlie many of the disputes that Carlson describes reflected strong philosophical biases that separated the more holistic approach found among European (especially German) geneticists, such as Richard Goldschmidt (who for a major figure gets scant treatment), from the reductionist approach of the American school. National distinctions such as those described so well by Jonathan Harwood in *Styles of Scientific Thought* (1993) do not form a significant part of Carlson's story. Historians of science will also look askance at the sharp distinction that Carlson draws between science that he claims lacks political content (such as the function of the accessory chromosome) from that which is overtly political (notably eugenics).

All in all, however, these are minor deficiencies in a book that will certainly have broad appeal, especially within the biological community. ■

Garland E. Allen is professor of biology at Washington University, St Louis, Missouri, USA.

A window on early animal evolution

The Cambrian Fossils of Chengjiang, China: The Flowering of Early Animal Life

by Xian-Guang Hou, Richard J. Aldridge, Jan Bergström, David J. Siveter, Derek J. Siveter & Xiang-Hong Feng
Blackwell: 2004. 248 pp. \$99.95, £60.

Zhe-Xi Luo

In the Early Cambrian period (544 million to 520 million years ago) there was a great evolutionary radiation of animal life. Within a short geological time, the vast majority of modern animal phyla (groups defined by their distinctive body plans) appeared in the fossil record worldwide. These early fossils represent the roots for the tree of all modern animal life.

Some of the jewels in the crown of Early Cambrian fossils are embedded in the mudstones of Chengjiang in southern China, where tens of thousands of exceptionally preserved and soft-bodied fossils have been discovered during the past 20 years. More than 130 species of 13 animal phyla have been recognized, and the number of organisms that are new to science is still growing. Aptly subtitled *The Flowering of Early Animal Life*, this beautifully produced book by Xian-Guang Hou and colleagues is the best systematic compendium of the entire Chengjiang biota, offering a rare view of this great episode in the diversification of animal life.

The book recounts the serendipitous discovery of this Lagerstätte, or great mother lode, of soft-bodied fossils in 1984. It also describes the persistent exploration over the next two decades by Hou, Jun-Yuan Chen, De-Gan Shu and their many collaborators that revealed the wealth of this site for studies of the early evolution of animal life. Several organisms from Chengjiang are so extraordinary that they cannot easily be placed in any modern animal phylum. For example, *Vetulicola* looks like a crustacean, judging by its tail, but has gill slits that some palaeontologists would argue indicate chordate affinities. These new fossils further widen the range of body plans for deuterostomes, a lineage that already includes echinoderms (sea urchins), various chordates and vertebrates.

These beautiful and unique fossils have inspired new scientific insights and led to the clashing of ideas. There is a great debate on the likely positions of Chengjiang animals such as the yunnanozoans in the deuterostome family tree. Such debates will surely redefine the phylogenetic framework for establishing the earliest evolution of key features of chordates.

The publication of this book is timely for several reasons. It is not only a fitting



Fossils found at Chengjiang provide evidence of the rapid evolution of a vast range of body plans.

Festschrift to celebrate the twentieth anniversary of Hou's discovery of the Chengjiang Lagerstätte, but it also provides an update on the fast-paced attempts to decipher the full evolutionary significance of this palaeontological treasure. It is also the first review in English on the Chengjiang biota. For palaeontologists and evolutionary biologists whose native tongue is not Chinese, this volume is a useful gateway to a much larger body of literature on the Chengjiang fossils, including some extensive Chinese monographs. For specialists, the book provides an extensive systematic survey of the Chengjiang organisms; each taxon has a synoptic review with key references. For generalists with an interest in early animal evolution, the book offers a concise overview of the geochronology, taphonomy, palaeo-environment and palaeoecology of the Chengjiang fossil sites. The book is lavishly illustrated with hundreds of colour photographs and accurate artistic rendering of many fossil organisms.

But beyond its utilitarian value as an informative reference resource, the book has much to offer those interested in a multidisciplinary approach to studying early animal evolution. The early animal fossil record, however incomplete, can tell us about the early diversification of major animal lineages, a hot topic for molecular evolutionary studies, especially with regard to the timing of early animal evolution. The Chengjiang fossils are the best source of evidence about the emergence of animal body plans, and have attracted interest from students of evolutionary development.

Many of my colleagues have proclaimed that the Chengjiang mudstone is China's answer to Canada's Burgess Shale. Notwithstanding my fondness for vertebrates, what makes Chengjiang special is the presence of new species of the earliest vertebrates and their putative chordate and deuterostome

relatives that are not known elsewhere; the Early Cambrian sea of Chengjiang really is a cradle of early chordate evolution. In such a context, this otherwise quite exhaustive book is a bit limited on what are Chengjiang's highlights: its vertebrates and their putative relatives. The authors cautiously avoided endorsing any of the competing interpretations of several enigmatic organisms, such as *Yunnanozoon*, almost to a fault.

The book's authors — six experts from five institutions in three countries — are a microcosm of the world community of palaeontologists who have long cherished collaborative research that transcends national boundaries. Such collaborations have accelerated the pace at which the major discoveries from Chengjiang are disseminated to the worldwide scientific community. This has not only won worldwide recognition for China's great fossil treasure, but has made the point that truly great science has no national boundaries. ■

Zhe-Xi Luo is a curator of vertebrate palaeontology, Carnegie Museum of Natural History, Pittsburgh, Pennsylvania 15213, USA.

Diagrams you can count on

Cogwheels of the Mind: the Story of Venn Diagrams

by A. W. F. Edwards (foreword by Ian Stewart)

Johns Hopkins University Press: 2004.
128 pp. \$25, £17

Jeremy Gray

Most of us know what a Venn diagram is, and many of us will have used them to represent families of objects with overlapping characteristics. Some of us will have

Science in culture

Return of the mummy

An ancient Egyptian priest has a virtual life at London's British Museum.

Michael Hopkin

Three thousand years ago, the people of Karnak, an ancient Egyptian community on the banks of the Nile, embalmed and buried a priest called Nesperennub. They did a good job — he has remained mummified ever since and is unlikely ever to be unwrapped. But now we can at last say hello, face to virtual face, thanks to a project that blends archaeology, medical technology and computer graphics.

The practice of unwrapping — and thus destroying — mummies was rife when Nesperennub was delivered into the collection of London's British Museum in 1899. But Nesperennub has survived to become the centrepiece of *Mummy: the Inside Story*, an exhibition that runs at the museum until January 2005. The exhibition showcases the museum's new non-invasive approach to studying wrapped mummies, using computerized tomography (CT) scanners of the sort common in hospitals.

John Taylor, the museum's Egyptologist who led the project, teamed up with the imaging company Silicon Graphics (SGI) to reconstruct Nesperennub from some 1,500 CT images, as shown here. This is the first time that CT scans have been used to reveal a mummy's intimate details — previous attempts using X-rays provided only sketchy images.

The new technique has sparked fresh speculation about Nesperennub's life and death, and confirms some long-held theories about the mummification process. The brain is missing, for example, as are the fine bones of the nasal cavity — a vindication of archaeologists' belief that the brain was removed by pulling it through the nose. Elsewhere, vital organs such as the lungs and liver have been removed, wrapped and replaced. Nesperennub's body is adorned with amulets, presumably to safeguard his passage to the world of the gods.

Judging by the wear and tear on his backbone, Nesperennub lived into his forties. And as his 1.62-metre stature attests, he was probably well fed. But life wasn't an endless round of pleasure: it seems he had a large abscess under one of his teeth; and a hole in his skull, apparently made from the inside, hints that he died of a brain tumour. The



project's members have even created a virtual bust of Nesperennub, although they admit that facial reconstructions always involve a measure of guesswork.

Most intriguingly of all, however, Nesperennub has a clay bowl seemingly stuck to the top of his head. Unlike the rest of his adornments, it is not ornately carved. Taylor speculates that it may have been placed there to catch drips of embalming resin and accidentally become stuck fast.

Visitors to the museum can ponder this possibility as they watch a film featuring computer reconstructions and scenes from everyday life in Karnak, as envisaged by the project team. Entertaining, yes. But Taylor emphasizes that the technique is first and foremost a research tool — the museum has about 100 mummies in its collection, and plans to bring many of them back to virtual life. Michael Hopkin writes for news@nature.com, Nature's online news service.

noticed that what works for three sets seems to work less well for four, and might break down entirely for five or more sets, when perhaps our enthusiasm for classification begins to fade. Is there an elegant way of drawing a Venn diagram showing all possible intersections of, say, seven sets, and is it systematic — does it build routinely to an arbitrary number of sets?

What *Cogwheels of the Mind* triumphantly

shows is that the answer to both questions is 'yes'. A.W. F. Edwards not only shows how it can be done, but discusses a variety of different solutions that connect Venn diagrams to significant questions in coding theory. He also provides a fascinating glimpse into how research into mathematics gets done, and tells a satisfying historical story. And he provides what may be the most important reason for continuing to care about these

diagrams: they illuminate the problem of finding maximal embeddings of subgraphs of a hypercube. All this and a great number of mathematically beautiful figures mean that the book deserves to become a minor classic and may well go on to make many friends for mathematics. ■

Jeremy Gray is at the Centre for the History of the Mathematical Sciences, Open University, Milton Keynes MK7 6AA, UK.

A radical reorientation

How an annotated book transformed a theoretician into an historian.

Owen Gingerich

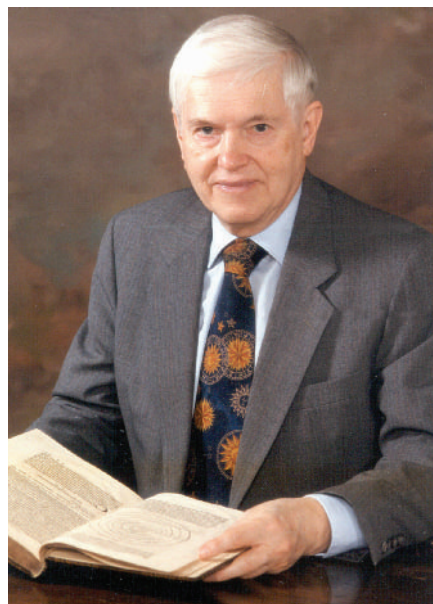
In 1970 I travelled to Cambridge, England, on a sabbatical leave, little dreaming that a serendipitous discovery was about to divert my career. The Harvard–Smithsonian reference model for the solar atmosphere had just been published. This was the culmination of a decade of computer work modelling the flow of radiation through stellar atmospheres and exploiting the latest observations made above the Earth's atmosphere by balloons, rockets and satellites. This paper, on which I was lead author, was destined to become a citation classic, and was to be the last straight astrophysics article I ever published. Although my goal — to understand better how science actually works — remained unchanged, an unanticipated turning point radically altered the grist for my mill.

Like many astronomers, I was looking forward to the 1973 quincentennial of the birth of Nicolaus Copernicus, the sixteenth-century cosmologist who proposed the heliocentric system. I was aware that the novelist Arthur Koestler, in his *The Sleepwalkers*, had taken Copernicus as his anti-hero and had declared that *De revolutionibus orbium coelestium* was “the book that nobody read” and “an all-time worst-seller”. Granted, there are probably more readers alive today than existed in the entire sixteenth century, and Copernicus' treatise is certainly formidably dense and highly mathematical. But was it really true that it had at best only a handful of readers?

In November 1970, I took my family to Scotland for a short holiday and en route I stopped off to discuss Copernicus with a fellow member of the committee planning the international quincentennial celebrations. We pondered the question of the Polish astronomer's readership, and came up with fewer than a dozen names of early astronomers who might have read a major part of it.

Two days later, at the Royal Observatory, Edinburgh, I stumbled on a first-edition *De revolutionibus* that had been brilliantly annotated from beginning to end. Something seemed curiously improbable. If the book had so few readers, how did it happen that the very next copy of Copernicus' text that I chanced to examine gave evidence of such thorough readership?

Eventually, I was able to identify the handwriting; the extensive marginalia were written at Wittenberg in the sixteenth century by Erasmus Reinhold, the leading astronomy teacher of the generation following



Journey through time and space: Owen Gingerich with his second-edition copy of *De revolutionibus*.

Copernicus, and one of the names on our shortlist of possible readers.

Undoubtedly, the most fascinating feature of Reinhold's annotations was the motto that he inscribed on the title page: “The Axiom of Astronomy: celestial motions are uniform and circular, or composed of uniform and circular parts.” To modern eyes, what a strange summary! Needless to say, no twenty-first-century astronomer would neglect to mention that Copernicus stopped the Sun and made the Earth into a planet, the essence of heliocentrism and the reason why today a first-edition *De revolutionibus* is a million-dollar icon.

What was going on here? Note that Copernicus had two independent aesthetic ideas, “theories pleasing to the mind” as he called them, for which he had sound philosophical grounds, but no observational proof at all. First was his well-known Sun-centred layout for the planetary system, literally establishing the Solar System. Second was his desire to achieve the required non-uniform orbital motion of the planets by combinations of uniform circular motions, something that appealed greatly to the model-building intentions of sixteenth-century astronomers, but which ultimately proved a dead end. Clearly, Reinhold appreciated the second aesthetic goal, but had understandable reservations about the first. The radical heliocentric cosmology threw the Earth into a dizzying motion that seemed ludicrously contrary to common sense, unsupported by any

physics and opposed to the traditional interpretations of holy scripture. The annotations within the book confirmed this stance: remarks on the cosmology were sparse, those on the technical details of the circles copious.

Reinhold's reaction helped solve a long-standing puzzle concerning the very gradual acceptance of the heliocentric blueprint, and it casts light on the reasons why Koestler went astray. With hindsight, we might have expected that astronomers would have immediately adopted Copernicus' vision of the Solar System, as for us it is so obviously the correct arrangement. As post-newtonians, we have no trouble envisioning the Earth as just one of the planetary family. Because the general acceptance of the new cosmology took place only several generations after *De revolutionibus* was published, Koestler must have jumped to the conclusion that no one had read the book.

The insights afforded by Reinhold's pattern of annotations set me on the trail of further copies of *De revolutionibus* to see if the book had other readers. What started as a simple research project turned into a fascinating, thirty-year obsession. Studying this turning point in the history of cosmology unexpectedly took my research on the nature of scientific discovery down a very different path.

One obvious conclusion is that Koestler had been dead wrong in saying that it was the book nobody read. *De revolutionibus* not only had an impressive roster of owners, but many of them read and annotated their copies — however, most of them were reading the book as a manual of geometrical model building, not as a physical description of the cosmos.

With the perspective of history, we can see that a persuasive coherency can finally displace an entrenched world view, but it takes time to build a constituency. The tortoise-like pace of the copernican revolution reflects the radical reorientation of thought required to accept that his “theory pleasing to the mind” described a real, physical universe and was not simply an imaginary device for calculating the positions of the planets. ■

Owen Gingerich is at the Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. A fuller account of his copernican adventures appears in *The Book Nobody Read: Chasing the Revolutions of Nicolaus Copernicus* (Walker, 2004), published in the UK in August (Heinemann, 2004).

Deeper understanding

Thomas S. Duffy

The boundary between the core and mantle is one of the most enigmatic regions of Earth's interior. Analyses of a newly discovered crystalline phase should yield a fuller understanding of this region.

The lowermost 250 km or so of Earth's mantle, known for historical reasons as D'', is comparatively small in volume but potentially holds the key to understanding a host of geophysical phenomena — among them the formation of plumes in the mantle, interactions between core and mantle, and the ultimate fate of subducting slabs of crust that are driven into the interior by tectonic forces. Investigations of this region largely depend on interpreting the behaviour of seismic waves, which have shown that it is highly complex. Until recently, however, studies of the region's mineral properties at high pressures and temperatures had been unable to provide satisfying explanations for much of this complexity. Part of the problem is that the extreme conditions in D'' — pressures up to 135 gigapascals and temperatures probably ranging between 2,000 K and 4,000 K — are difficult to reach in the laboratory. However, laboratory experiments and theory are finally coming together to bring this region into sharper focus.

Beginning on page 442 of this issue, papers by Iitaka *et al.*¹ and Oganov and Ono² provide insights that link the calculated physical properties of a newly discovered³ high-pressure crystal structure with seismic observations of the deep lower mantle. Earth's mantle is composed mostly of dense silicate minerals containing magnesium, iron, calcium and aluminium. Experiments have shown that the lower mantle, extending from 660 km depth to the base of the mantle at about 2,900 km (Fig. 1), is mainly composed of (Mg,Fe)SiO₃ in a crystal structure known as perovskite. Although the properties of this material are compatible with most observations for the lower mantle, the abrupt change in properties near the mantle's base defied explanation in terms of perovskite behaviour.

Thus, Murakami and colleagues' experimental discovery³ of a 'post-perovskite phase' in MgSiO₃, at conditions comparable to the D'' region, has stimulated considerable interest in the physical properties of the new phase. Given the difficulty of performing direct experiments under these conditions, first-principles quantum mechanical calculations of the type carried out by Iitaka *et al.*¹ and Oganov and Ono² are especially useful for studying the deep Earth. In contrast to the perovskite structure that is widely adopted by many compounds, the

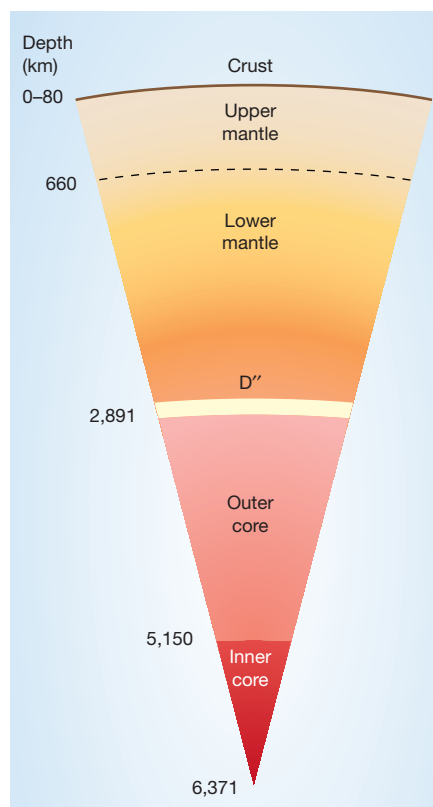


Figure 1 The principal regions of Earth's interior. The latest interpretations of the lowermost mantle, the D'' layer, are shown in Fig. 2.

post-perovskite phase seems to be rather uncommon. In this structure, each silicon cation remains surrounded by six oxygen anions — producing the octahedral coordination that is characteristic of the lower mantle. But rather than forming a corner-linked, three-dimensional network as in perovskite, in the post-perovskite phase the silicon octahedra share edges and corners to form a sheet-like structure with alternating magnesium and silicon layers (see Fig. 1 of Iitaka and colleagues' paper on page 442).

This much could be demonstrated by laboratory experiment⁴. But theoretical calculations now shed light on several other properties of the post-perovskite phase. First, by taking account of thermodynamic considerations, Iitaka *et al.*¹ and Oganov and Ono² show that it is expected to be stable at pressures above about 100 GPa (at 0 K). Oganov and Ono² also include experimental observations of the new phase after heating at near 118 GPa. For experiments under

these extreme conditions, reports of structural changes are all too frequently not verified. So multiple experimental observations of the post-perovskite phase²⁻⁴, together with the theoretical predictions of its stability, mean that the implications of the new phase need to be seriously considered. Indeed, the two theoretical reports here^{1,2} are also generally compatible with the findings from another study^{5,6}.

In conjunction with experimental findings, the theoretical results at 0 K indicate that the transition has a positive pressure-temperature slope. At mantle temperatures, the phase transition is then expected to occur about 200–300 km above the base of the mantle (Fig. 2, overleaf), consistent with evidence for a sudden change, or discontinuity, in the velocity of seismic waves there^{7,8}, possibly global in extent^{9,10}. The positive slope of the phase boundary is even compatible with seismic-wave evidence¹⁰ that the D'' discontinuity is elevated in seismically fast (and presumably cold) regions and depressed in seismically slow (hot) areas. The new phase is also found to be about 1–2% denser than perovskite at D'' conditions.

Calculations of the elastic properties confirm that the post-perovskite phase is anisotropic, being more compressible normal to layering than parallel to it. It also exhibits considerable anisotropy in seismic-wave velocities, especially for the type of waves known as shear waves. Given the nature of the structure, it is plausible that post-perovskite crystals will develop a lattice-preferred orientation under compression such that the direction normal to layering will tend to lie along the vertical. If the layering is imperfect, and taking into account the presence of other phases, the calculations show that a 2–3% seismic discontinuity consistent with deep-mantle observations⁷⁻¹⁰ would result from this phase transition. The transition is expected to produce a larger discontinuity for shear waves than for compressional waves, the other main type of seismic wave.

This form of texturing will also result in horizontally polarized shear waves (of velocity v_{SH}) propagating faster than vertically polarized shear waves (v_{SV}). Seismic anisotropy in D'' is complex, but this sense of anisotropy has been well documented in certain regions¹¹. Previously, this behaviour was difficult to reconcile with the known elastic

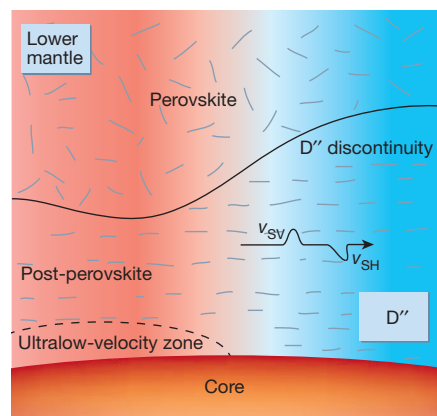


Figure 2 New model for the mantle's base. The D'' discontinuity is now thought to be due to a transition from the perovskite to a post-perovskite structure in (Mg,Fe)SiO₃ about 200–300 km above the base of the mantle. The phase boundary is elevated in locally cooler regions (blue) and depressed in locally hotter regions (red). A tendency for the layers of the post-perovskite phase to align parallel to Earth's core can help to explain the faster propagation of horizontally polarized (v_{SH}) than vertically polarized (v_{SV}) shear waves. Ultralow-velocity zones are thin (5–40-km thick) regions, located directly above the core, where shear-wave velocities are strongly depressed.

properties and deformation behaviour of perovskite and other lower-mantle minerals. Instead, it was proposed that the anisotropy resulted from aligned inclusions or layering of minerals with dissimilar seismic velocities. The discovery of the post-perovskite phase may provide a simpler explanation.

The proposed transition between perovskite and post-perovskite will not resolve all questions about the D'' region. But it clearly provides a new framework for studying the region and is sure to stimulate further geophysical observations, laboratory experiments and computer calculations. From a mineral-physics viewpoint, studies of texture development in the new phase, as well as constraints on the behaviour of more chemically complex systems, are clearly needed. Also, the elastic anisotropy has only been calculated at 0 K, yet in some cases temperature can drastically change the magnitude and even orientation of anisotropy. The theoretical studies^{1,2,5,6} are in remarkably good agreement. But they all used similar techniques involving some degree of approximation, which will also necessitate further examination.

Nevertheless, a new era in the study of Earth's deepest mantle has begun. An explanation for both the D'' discontinuity and the onset of seismic anisotropy in the region may finally be within our grasp. ■

Thomas S. Duffy is in the Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA.
e-mail: duffy@princeton.edu

1. Iitaka, T., Hirose, K., Kawamura, K. & Murakami, M. *Nature* **430**, 442–445 (2004).
2. Oganov, A. R. & Ono, S. *Nature* **430**, 445–448 (2004).
3. Murakami, M., Hirose, K., Kawamura, K., Sata, N. & Ohishi, Y. *Science* **304**, 855–858 (2004).
4. Shim, S.-H., Duffy, T. S., Jeanloz, R. & Shen, G. *Geophys. Res. Lett.* **31**, L10603 (2004).
5. Tsuchiya, T., Tsuchiya, J., Umemoto, K. & Wentzcovitch, R. M. *Earth Planet. Sci. Lett.* **224**, 241–248 (2004).
6. Tsuchiya, T., Tsuchiya, J., Umemoto, K. & Wentzcovitch, R. M. *Geophys. Res. Lett.* **31**, doi:10.1029/2004GL020278 (2004).
7. Lay, T. & Helmberger, D. V. *Geophys. J. R. Astron. Soc.* **75**, 799–837 (1983).
8. Wyssession, M. E. *et al.* in *The Core–Mantle Boundary Region* (eds Gurnis, M. *et al.*) 273–297 (Am. Geophys. Un., Washington DC, 1998).
9. Nataf, H.-C. & Houard, S. *Geophys. Res. Lett.* **20**, 2371–2374 (1993).
10. Sidorin, I., Gurnis, M. & Helmberger, D. V. *Science* **286**, 1326–1331 (1999).
11. Kendall, J.-M. & Silver, P. G. *Nature* **381**, 409–412 (1996).

Language

Children think before they speak

Paul Bloom

A linguistic contrast between English and Korean provides a telling test of different ideas about whether thought precedes the acquisition of language, or whether certain concepts are language-specific.

In his autobiography, written in the fourth century AD, Saint Augustine¹ described how he learned to talk: “By constantly hearing words, as they occurred in various sentences, I collected gradually for what they stood, and having broken in my mouth to these signs, I thereby gave utterance to my will.” For Augustine, thought precedes language: language is a tool with which to express one's ideas and to understand the ideas of others. This is the view of many contemporary philosophers and psychologists^{2–4}, but it is not the only possibility. Many scholars would instead endorse the theory of linguistic relativity, and maintain that learning a language has a profound influence on a child's mental life. If so, then speakers of different languages might think in very different ways^{5,6}.

On page 453 of this issue, Hespos and Spelke⁷ present data, from 5-month-old babies, that support Saint Augustine's view. They concentrate on a much-studied linguistic distinction between ‘tight-fitting contact’ and ‘loose-fitting contact’. For instance, Korean uses different verbs when describing placing a shoe in a large box, where it fits loosely, and when placing the shoe in a small box, where it is a tight fit — even young children who are just beginning to learn Korean honour this distinction when they speak⁶. Hespos and Spelke ask whether this distinction between two sorts of contact is universal, and exists before language-learning (in which case it should be present in babies), or whether it is the result of acquiring Korean (in which case it should be present only in children and adults who have some knowledge of that language).

They address this question by using a standard method in infant cognition. They show babies instances of a given category until they get bored (or habituated) and stop looking, and then see if the babies perk up — look for longer — at an instance from a



Snug fit: in situations such as this, the Korean language makes a distinction between tight and loose contact.

new category. If so, it means that babies are sensitive to the categorical difference. Using this method, Hespos and Spelke find that 5-month-olds who are raised in an English-speaking community are sensitive to the Korean categories of meaning. If the babies are habituated to tight-fitting events, such as a cylinder placed within a narrow container or a ring-like object placed around a post, they will look for longer when later shown a loose-fitting event, such as a cylinder placed into a wide container (see Figs 1 and 2 of the paper, pages 453 and 454). The converse is also true. If habituated to loose-fitting events, babies will look for longer when shown a tight-fitting event. In this domain at least, the traditional view is right: thought precedes language.

Hespos and Spelke note the analogy here with phonology, in which there are also cross-

S. OLIVER/DK IMAGES

linguistic differences — certain acoustic contrasts are present in some languages but not others. It is not that children become increasingly sensitive to the distinctions made in the language that they are exposed to. Instead, they start off sensitive to every distinction that human languages make; the process of learning a particular language involves becoming insensitive to those distinctions that are irrelevant, and learning what to ignore.

Phonology and meaning differ in certain important ways, however. Another striking fact about phonological development is that the early sensitivity disappears. If the child's language does not exploit a distinction, then the child loses the ability to notice it. This is one reason why it is so difficult to learn a second language. But, as Hesplos and Spelke point out, even an adult English-speaker who has never heard Korean can tell the difference between a tight fit and a loose fit. This difference between phonology and meaning makes sense. Phonology is for communicating; once a language is learned, nothing is lost by jettisoning those phonological contrasts that are irrelevant. But meaningful contrasts such as loose fit and tight fit are for making sense of the world. This is nicely demonstrated by the finding that 5-month-olds can use their sensitivity in a non-linguistic context, when predicting the motions of objects.

In addition, although all phonological distinctions made by language may be innate, this cannot be true for all distinctions of meaning. Babies might understand the contrast between tight fit and loose fit, and between support and containment, but they are unlikely to comprehend the contrasting meanings of the verbs 'leering' and 'glaring', or the nouns 'accountant' and 'lawyer'. This must be learned.

What is the nature of this learning? One compromise view is that there is a universal core of meaningful distinctions that all humans share, but other distinctions of meaning that people make are shaped by the forces of language; this is consistent with the theory of linguistic relativity. But it is also possible that the strong Augustinian view is correct: language learning might really be the act of learning to express ideas that already exist, either because they are unlearned (as is likely to be true of the domain studied by Hesplos and Spelke) or because they have been learned though experience with the physical and social world.

The question of how language and thought are related is one of the deepest in psychology, and there are many variants of the claim of linguistic relativity that this current research does not address^{8–10}. But the capacities of 5-month-olds do pose a serious challenge for certain strong versions of the view that language precedes thought, and show that, in some domains at least, children think before they speak. ■

Paul Bloom is in the Department of Psychology,

Yale University, PO Box 208205, New Haven, Connecticut 06520-8205, USA.
e-mail: paul.bloom@yale.edu

1. Saint Augustine *The Confessions of Saint Augustine* (Random House, New York, 1961).
2. Fodor, J. A. *The Language of Thought* (Crowell, New York, 1975).
3. Bloom, P. *How Children Learn the Meanings of Words* (MIT Press, Cambridge, Massachusetts, 2000).

Palaeontology

Echinoderm roots

Andrew B. Smith

A bold claim about the origins of the echinoderms is based on newly discovered fossils from China. But many pieces are still missing from this part of the fragmented puzzle of life's evolutionary history.

Few marine animals are so immediately recognizable as echinoderms. The five-fold symmetry of a starfish or sea urchin is striking (Fig. 1), and this pentaradial form sets them apart from their bilaterally symmetrical relatives. The echinoderm skeleton is equally distinctive, being made of calcite plates with a microstructure that resembles a very holey Swiss cheese. Finally there is their bizarre asymmetrical transformation from larva to adult, which involves loss of the right-hand set of paired larval body chambers. But echinoderms have not always possessed these features, and unravelling their early history remains highly controversial.

On page 422 of this issue Shu *et al.*¹ describe a new group of small fossils from the Lower Cambrian of Chengjiang, China. The fossils are some 520 million years old: Shu *et al.* call them vetulocystids, and interpret them as the most primitive echinoderms yet known. If correct, this links the echinoderms to an enigmatic group, the vetulicolians, remains of which are found in the same deposits of early Cambrian age.

Fossils can be made sense of only by comparison with living organisms, where the biological function of structures that become preserved can be directly observed. This is relatively easy when the fossil is rather close in structure to its extant relatives. But fossils such as the vetulocystids, whose affinity with living groups is not immediately apparent, pose a major difficulty. In the past they would have simply been hived off into their own higher taxonomic group, thereby avoiding the problem. Nowadays, palaeontologists take the harder path and strive to place such fossils into their appropriate branch in the tree of life, interpreting characters they display in the most plausible (or least implausible) way.

The vetulocystids are a fascinating but frustrating group — fascinating, because of what they may tell us about the morphology of echinoderms before they had evolved such distinctive morphology, and frustrating because of the difficulty of interpreting even

4. Gleitman, L. R. *Language Acquisition* 1, 3–55 (1990).
5. Whorf, B. L. *Language, Thought, and Reality* (MIT Press, Cambridge, Massachusetts, 1956).
6. Choi, S. & Bowerman, M. *Cognition* 41, 83–121 (1991).
7. Hesplos, S. J. & Spelke, E. S. *Nature* 430, 453–456 (2004).
8. Xu, F. *Acta Psychol.* 102, 113–136 (1999).
9. Tager-Flusberg, H. & Sullivan, K. *Cognition* 76, 59–89 (2000).
10. Bloom, P. & Keil, F. *Mind Lang.* 16, 351–364 (2001).



Figure 1 Sea-urchin symmetry: an example of the unique pentaradial form of echinoderms.

their basic anatomical organization. In this they are not alone. Another group of primitive and entirely extinct echinoderms, the homalozoans, have been the source of debate amongst palaeontologists for years.

Echinoderms belong to the branch of the animal kingdom known as the deuterostomes, a group that also includes ourselves. This is a very diverse assemblage of organisms, the other members being the vertebrates (fishes and tetrapods), urochordates (tunicates and sea squirts) and hemichordates (pterobranchs and acorn worms) (Fig. 2, overleaf). Molecular data strongly support this grouping, placing vertebrates and urochordates together and echinoderms and hemichordates as a second pairing². But in terms of morphology echinoderms have always stood apart because of their aberrant symmetry and lack of structures known as gill slits. Gill slits are present in hemichordates, urochordates and the more primitive vertebrates, and are openings that pierce the wall of the digestive system just behind the mouth. They appear to have evolved for venting excess water drawn into the gut during feeding and are unique to deuterostomes.

Homalozoans are important fossils because they help bridge the gap between radiate echinoderms and other



100 YEARS AGO

The Fourth Dimension. By C. Howard Hinton, M.A. A book bearing the present title may be reasonably expected to contain certain things. In the first place it should have a clear exposition of Descartes's applications of algebra to geometry, and conversely of geometry to algebra, the logical conclusion of which consists in the removal of all restrictions as to the conceivable number of dimensions of space. In the second place it should contain clear, concise, and exactly worded statements of the peculiar and distinctive geometrical properties which are characteristic of spaces of two, three, four, or more dimensions respectively... Now such things as these are either entirely absent from the book or else they are mixed up with such a mass of irrelevant and discursive matter as to render it often quite impossible to make out what the author is driving at... There is a certain class of individual, far too common in this country, who busies himself in pestering his mathematical friends with long and rambling letters on such questions as "What is the fourth dimension?" or "What is the ether?" Such people very rarely know anything about the three dimensions of the space they live in, but Mr. Hinton's book will, it is to be hoped, give them something to think about which will at least amuse them and keep them occupied.

From *Nature* 21 July 1904.

50 YEARS AGO

It is firmly believed by most theoretical physicists that the antiparticle to a proton can exist... This particle should bear the same relation to a proton as does the positron to an electron. Should it be possible to create such a particle by a proton-nucleon collision, the incident proton will require a kinetic energy of at least 5.6×10^9 electron volts. At present only one particle accelerator, the Berkeley 'Bevatron', exists which is designed to reach this energy. At this energy the production of antiprotons is not expected to be very copious, and their abundance relative to other particles formed will probably be far smaller than in many cosmic ray phenomena which have already been extensively investigated. The problem of the conclusive identification of the antiproton may be a difficult one for some time to come.

From *Nature* 24 July 1954.

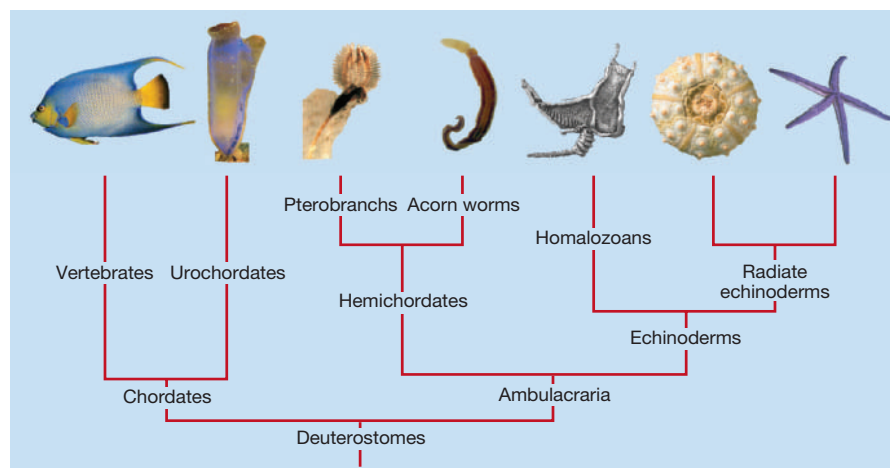


Figure 2 The main deuterostome groups. Shu *et al.*¹ interpret their vetulocystid fossils as an offshoot of the echinoderm lineage that branched off before the division between the now-extinct homalochoans and the radiate echinoderms. According to this view, the vetulocystids have affinities with another mysterious group, the vetulicolians, which may have diverged from the deuterostome line before the division between the chordates and ambulacraria.

deuterostomes. They have the characteristic skeleton of an echinoderm and a strongly asymmetrical development. But homalochoans never display even a hint of radiate symmetry and, more importantly, they have gill slits and a post-anal muscular appendage. These last two traits are probably primitive features common to all deuterostomes. So do the vetulocystids take us even further back towards the root of echinoderms?

Faced with such fossils, the first step is to define front and back, and identify the major body openings — mouth and anus, gill slits, and the pores involved in water circulation or reproduction (gonopores) are all possibilities that must be considered. Here the problems really start because fossils do not come ready labelled. Pictures of the fossils themselves, and the interpretations of Shu *et al.*, appear in Figs 1–3 of their paper (pages 423–425).

In the best-preserved vetulocystid material there is evidence for a straight gut running to the posterior along a possibly segmented tail-like structure, as in vetulicolians³, and this presumably led to a terminal anus. The sack-like anterior part, or theca, has two large and more or less identical openings, both circular and taking the form of a low pleated cone. The anterior of the two circular cones is taken to be the mouth and the posterior may be the anus or gonopore. Towards the base of the theca there is a rhomboidal structure, which Shu *et al.* interpret as respiratory in function and possibly a gill. The identification of the vetulocystids as a 'basal' echinoderm group hinges on this interpretation, with the presence of a single rather than paired 'gill' implying echinoderm-like developmental asymmetry. There is no calcite plating.

But other interpretations are possible. The two circular structures could be gill openings (they are in a similar position in vetulicolians³), the mouth anterior and forming the elongate recessed zone, and the

rhomboidal structure a folded epithelial zone for gaseous exchange (analogous to those of certain primitive stemmed echinoderms) rather than a gill. That would make the echinoderm affinities of vetulocystids much more dubious. Vetulocystids and vetulicolians share many similarities and are clearly closely related, but quite where they fit within the deuterostomes remains ambiguous because of these alternative possibilities. Such is the way with fossils.

There is now direct fossil evidence that all of the major deuterostome groups were established by about 520 million years ago. Fossil vertebrates (yunnanozoans⁴), tunicates (*Shankouclava*⁵) and both asymmetric and radiate echinoderms (homalochoans, helicoplacoids) have all now been discovered in early Cambrian deposits. *Phlogites*, a tentacle-bearing early Cambrian fossil of uncertain affinity⁵, might even be a hemichordate or part of the common ancestral lineage of echinoderms and hemichordates. So, if deuterostome divergence occurred around 575 million years ago, as recent molecular-clock studies suggest⁶, there is a 50-million-year gap in the fossil record between the origin of deuterostomes and their appearance in the fossil record. In the jigsaw of deuterostome evolution, vetulocystids represent another piece to be fitted into a puzzle where many of the pieces are still missing. ■

Andrew B. Smith is in the Department of Palaeontology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.
e-mail: abs@nhm.ac.uk

1. Shu, D.-G., Conway Morris, S., Han, J., Zhang, Z.-F. & Liu, J.-N. *Nature* **430**, 422–428 (2004).
2. Smith, A. B. *et al.* in *Assembling the Tree of Life* (eds Cracraft, J. & Donoghue, M. J.) Ch. 22 (Oxford Univ. Press, 2004).
3. Shu, D.-G. *et al.* *Nature* **414**, 419–424 (2001).
4. Mallatt, J. & Chen, J.-Y. *J. Morphol.* **258**, 1–31 (2003).
5. Chen, J.-Y. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 8314–8318 (2003).
6. Peterson, K. J. *et al.* *Proc. Natl Acad. Sci. USA* **101**, 6536–6541 (2004).

Liquid crystals

A missing phase found at last?

Geoffrey R. Luckhurst

Molecules that form liquid crystals are usually rod-like, but bend them and a new liquid-crystal phase — a biaxial nematic — should form. Strong evidence for the existence of this phase has only now emerged.

Liquid crystals form the basis of several flat-screen display technologies. The alignment of their rod-like molecules, in what is known as the 'nematic' phase, can be achieved rapidly and with low voltage. According to theory, these materials should also exist in another nematic phase with potential for use in display applications. In *Physical Review Letters*, Madsen *et al.*¹ and Acharya *et al.*² report firm evidence for this so-called biaxial nematic phase. The announcement has created considerable excitement, for it opens up new areas of both fundamental and applied research. It seems that a Holy Grail of liquid-crystal science has at last been found.

The hunt for this new liquid-crystal phase began more than 30 years ago, when it was recognized that the molecules forming liquid crystals deviate from their presumed cylindrical shape³. In fact, the molecules are more lozenge-like, and it is because of this lowering of the molecular symmetry that two nematic phases should be possible. In both phases, the molecular axes tend to align over large distances, although their centres of mass have no long-range order. In the common nematic phase, the long axes of the molecules tend to be aligned, in a direction known as the 'director', but the short axes are not ordered (Fig. 1a). This is known as the

uniaxial nematic, because its properties are cylindrically symmetric about the director. The other nematic phase is predicted to occur at lower temperatures than the uniaxial nematic. In this instance, the molecular short axes tend to be aligned over large distances as well, giving a phase with three directors, about which three molecular axes tend to align. The phase has two optic axes (directions along which the optical properties appear to be cylindrically symmetric), hence its description as 'biaxial'.

The design of molecules with sufficient biaxiality is a difficult task. There are many forms that the molecules could take, in addition to the lozenge, that deviate from a cylindrical shape. For example, a series of molecular shapes could be formed by linking rod-like, disc-like and semicircular moieties together⁴ (Fig. 1b). Such molecules would be long enough to exhibit a nematic phase and to deviate sufficiently from cylindrical symmetry to give a biaxial nematic. Indeed, the first claimed discovery of a biaxial nematic⁵ was for a compound formed of spoon-like molecules, and similar claims soon followed for cross-shaped⁶ and bone-shaped⁷ molecules.

Although such molecules should certainly exhibit biaxial nematics, the problem lies in identifying the symmetry of the nematic

phase unambiguously⁸. It is now recognized that optical techniques commonly used to investigate other liquid-crystal phases may incorrectly identify a uniaxial nematic as biaxial. NMR spectroscopy seems to provide one of the most definitive tests. When applied to materials purportedly forming biaxial nematics, it has always shown the phase to be uniaxial⁹. This disappointing result indicates just how exacting the design criterion must be if the transition to the biaxial nematic is not to be blocked by the material freezing first. It is also frustrating, given that biaxial nematic phases are already known to form in classes of liquid crystal other than that used for display technologies (such as lyotropic¹⁰ and polymeric¹¹ systems).

One design not previously explored was that of a V-shaped molecule, formed simply by bending a rod-like molecule. Such boomerang-like molecules should form a biaxial nematic^{9,12} (Fig. 1c) and, more importantly, theory has indicated how the stability of the elusive phase should depend on the angle between the two arms. Researchers have been synthesizing materials with bent molecules, and some had begun to suspect that the nematic phases formed were biaxial¹³. However, the definitive NMR experiments proved especially challenging and have only now been performed¹. Although some of Madsen and colleagues' results¹ are puzzling, others do indicate a small biaxiality for the nematic phase, in agreement with optical results. This identification is consistent with Acharya and colleagues' X-ray diffraction studies², in which the minor director was aligned by applying an electric field.

The discovery has important implications. For example, it raises the question of the mechanism underlying the formation of the biaxial nematic. So far, this has been attributed to molecular shape. But for these boomerang molecules this seems unlikely, because their apex angle of around 140° is far from the predicted optimum value^{9,12} of 109°. Indeed, Madsen *et al.*¹ recognize this and have suggested that strong electrostatic forces are crucial, at least for these materials, in stabilizing the biaxial nematic. This would be especially intriguing, for it could result in a nematic with ferroelectric properties.

What of the future? Clearly, theoreticians will want to think again about the molecular design of biaxial nematics. No doubt, given the controversial nature of previous claims, others will seek to test this latest subtle identification of the biaxial nematic phase. More importantly, the discovery will stimulate the search for other examples of biaxial nematics, especially those formed by V-shaped molecules. In fact, many such materials are already available, although produced for other reasons, and would certainly merit re-examination. The fundamental behaviour of

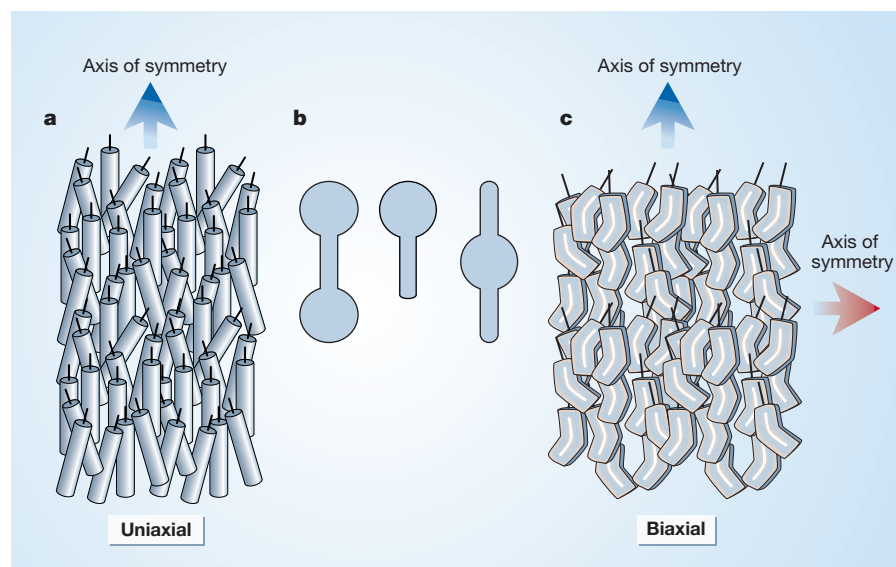


Figure 1 Nematic phases. a, Rod-like molecules can organize themselves so that their long axes are aligned. This is the uniaxial nematic phase, characterized by a single optic axis, known as the director. b, For molecules whose shape deviates from the cylindrical form, as outlined here, a second nematic phase should exist. This is the biaxial nematic phase, with two optic axes. c, This phase now seems to have been found for boomerang-shaped molecules^{1,2}.

this phase, both static and dynamic, must be explored, creating a new area of research in the field of liquid-crystal science and technology.

Geoffrey R. Luckhurst is in the School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

e-mail: gl@soton.ac.uk

- Madsen, L. A., Dingemans, T. J., Nakata, M. & Samulski, E. T. *Phys. Rev. Lett.* **92**, 145505 (2004).
- Acharya, B. R., Primak, A. & Kumar, S. *Phys. Rev. Lett.* **92**, 145506 (2004).
- Freiser, M. J. *Phys. Rev. Lett.* **24**, 1041–1043 (1970).

- Praefcke, K. *et al.* *Mol. Cryst. Liq. Cryst.* **198**, 393–405 (1991).
- Malthête, J., Nguyen, H. T. & Levelut, A. M. *J. Chem. Soc. Chem. Commun.* 1548–1549 (1986); Corrigendum 40 (1987).
- Chandrasekhar, S., Ratna, B. R., Sadashiva, B. K. & Raja, N. V. *Mol. Cryst. Liq. Cryst.* **165**, 123–130 (1988).
- Praefcke, K. *et al.* *Liq. Cryst.* **7**, 589–594 (1990).
- Galerie, Y. *Mol. Cryst. Liq. Cryst.* **323**, 211–229 (1998).
- Luckhurst, G. R. *Thin Solid Films* **393**, 40–52 (2001).
- Yu, L. J. & Saupe, A. *Phys. Rev. Lett.* **45**, 1000–1003 (1980).
- Severing, K. & Saalwächter, K. *Phys. Rev. Lett.* **92**, 125501 (2004).
- Teixeira, P. I. C., Masters, A. J. & Mulder, B. M. *Mol. Cryst. Liq. Cryst.* **323**, 167–189 (1998).
- Dingemans, T. J. & Samulski, E. T. *Liq. Cryst.* **27**, 131–136 (2000).

Animal behaviour

A social call

Christopher B. Sturdy

An indicator of animal intelligence is thought to be the ability to judge relationships between members of the same species. This talent, previously seen only in primates, seems to be evident in a bird.

Zebra finches are small, colonial songbirds native to Australia, and they have been favourite subjects of biologists for years¹. Most studies have focused on the birds' vocal-communication behaviour, and many insights have emerged from such research. For example, female zebra finches can recognize their mate, with whom they form lasting pair bonds, by his 'distance call', one of the most frequent vocalizations of these finches². But the early studies found no reciprocal recognition — male zebra finches did not recognize their mate through female distance calls. The reason for this was unclear, although it was speculated that, as the females' calls have a simpler acoustic structure than those of the males, the calls contained fewer individually recognizable features.

One overlooked factor was the context in which these recognition tests were conducted. The birds, normally members of large flocks in the wild or large groups in captivity, had been tested while alone. On page 448 of this issue, Vignal *et al.*³ suggest not only that female zebra finch distance calls are individualized, but that social isolation in the previous tests was the primary cause of the male's apparent recognition failure. They go a step further, showing that, remarkably, the males respond to their mate's calls according to the social situation. This latter finding suggests for the first time that a non-primate may be able to assess the social relationships between other animals of its own



In mixed company: Vignal *et al.*³ find that male zebra finches (with orange cheek patches and chestnut flanks) respond to their mate's call in a manner that depends on the social context.

species, an ability thought to be a mark of intelligence.

To determine whether the females' distance calls had the potential for individual recognition, Vignal *et al.*³ analysed the acoustic structure of calls from seven female zebra finches. Several features were measured to characterize each female's call. By quantifying the variability of each acoustic feature in each bird and comparing these with the variability found for the group, Vignal *et al.* concluded that certain features were sufficiently individualized to allow recognition. Further, when these features were analysed in more detail, the authors found that they could identify individual females with 100% accuracy. Clearly, males should be able to identify their mate's calls. Why, then, was this not seen before? And what might facilitate this recognition process in male zebra finches? The answer seems to lie in the birds' brains.

Two neurobiological studies^{4,5} of brain

activation during singing in zebra finches found that the areas of the brain responsible for song production and for perception are differentially activated according to the social context in which the bird is placed. Both *in vivo* electrophysiological recordings and assessment of early gene activity indicated that different brain areas are activated depending on whether the males are singing alone or with a female. As all previous studies of mate recognition by male zebra finches were conducted when the bird was alone, could it be that differences in social context modify recognition behaviour in an analogous way to the modification of brain-activation patterns?

To address this question directly, Vignal *et al.*³ played back either the mate's call or a familiar female's call to male zebra finches in three different 'social' conditions: with two unmated males, with a mated pair or with an unmated pair. The results were striking. When male zebra finches were with either unmated males or an unmated pair, they showed no greater response to their mate's call than to the call of a familiar female. That is, males called back at about equal rates in response to both their mate's calls and those of the familiar female. But when males were exposed to calls in the presence of a mated pair, the differences were marked: males called back almost twice as often in response to their mate's calls as to the familiar female's calls.

These results are intriguing on several levels, and will undoubtedly provide fodder for future research. First, Vignal *et al.* show that even a seemingly simple call can contain enough information to be identified individually. Equally interesting is how this comes about in female zebra finches, which until now had been thought to be unable to learn vocalizations like males of the species. Further, how do social context and the male's recognition ability interact? Do males recognize their mates in all social situations but express this ability only in certain ones? Or does the social situation itself activate the brain areas required for recognition, facilitating the recognition process?

Perhaps the most important implication of this study is that we primates are not alone in our ability to judge social context. It seems that birds are far more aware of such niceties than we have previously given them credit for.

Christopher B. Sturdy is in the Department of Psychology, Centre for Neuroscience, University of Alberta, Edmonton AB T6G 2E9, Canada.

e-mail: csturdy@ualberta.ca

- Zann, R. *The Zebra Finch: A Synthesis of Field and Laboratory Studies* (Oxford Univ. Press, 1996).
- Miller, D. B. *Anim. Behav.* **27**, 376–380 (1979).
- Vignal, C., Mathevon, N. & Mottin, S. *Nature* **430**, 448–451 (2004).
- Hessler, N. A. & Doupe, A. J. *Nature Neurosci.* **2**, 209–211 (1999).
- Jarvis, E. D., Scharff, C. & Grossman, M. R. *Neuron* **21**, 775–788 (1998).

Obituary

Thomas Gold (1920–2004)

Thomas Gold, who died on 22 June, was born in Vienna, Austria, emigrating to Britain along with his family in 1933. Following school in Switzerland, in 1939 he started studying mechanical sciences at Trinity College, University of Cambridge.

In his youth Gold never envisaged becoming a scientist, surprisingly so given the imaginativeness and variety of his later contributions. Evidently it was his experiences during the Second World War that brought science to the forefront of his thinking. He and I lived together in 1940 (in a British internment camp) and again (voluntarily) in 1942–45. In 1942, first I, then Gold, joined Fred Hoyle at the naval radar research establishment, working for the British Admiralty. From 1943, Hoyle, Gold and I spent many evenings together, and Hoyle's enthusiasm for astrophysics soon infected us both.

After the war, again in Cambridge, the three of us resumed our conversations. One of our subjects was cosmology. The main difficulty then was the discrepancy between the timescale of our neighbourhood (the ages of the Earth and the Sun) and that of the rate at which distant galaxies are receding, as determined by Edwin Hubble. This was so serious that such distinguished scientists as Paul Dirac, Edward Milne and Arthur Eddington proposed theories involving major secular changes to the laws of physics. This made us feel that one should first consider a stationary Universe. But the observed Hubble expansion was incompatible with the constant density of matter required for a stationary Universe.

We were stumped over this — until Gold proposed the idea of the continual creation of matter in space. We found no contradictions with known observations. Hoyle chose a different route from Gold and me, so the 'steady-state theory of the expanding Universe' was presented in two separate papers in the autumn of 1948. But only a limited number of astronomers found the theory attractive, and a few were hostile.

Early attempts to disprove our theory failed. Later, developments in astronomical instruments made the subject much more complicated. Most scientists now consider our theory to be disproved, but Gold thought to the end that it was the only valid analysis in cosmology and that a proper critique of the observations would remove any contradictions. Hoyle and two colleagues modified the theory, making it compatible with all observations. But they abandoned the stationary character of our



A singular scientist

model and therefore its chief attraction for Gold and me.

My interests have changed, and I have not worked in cosmology for half a century, so Gold's stand was a lonely one. Unfortunately, there is nobody knowledgeable in modern astronomy who is willing and able to carry out the critique he had hoped for.

Gold also studied questions involving magnetism in the Earth and in space. He and I wrote three joint papers, the last one dealing with the theoretical problem of radiation from a uniformly accelerated electric charge. But he soon moved on to thinking about human hearing. This was characteristic: he would see a difficulty in an unfamiliar subject and become intent on resolving it. The excellent frequency discrimination of our sense of hearing could not then be explained. Gold saw the similarity of the problem to that of receiver design for radio, radar and television, where the sensitivity has to be restricted to a narrow band of frequencies to avoid 'crosstalk'. This is achieved by the use of positive feedback. Gold proposed that the same principle applies to the human ear. This brilliant contribution led to his election to a fellowship at Trinity, but only some 30 years later was his theory accepted by physiologists.

From Cambridge, Gold went to the Royal Greenwich Observatory for the

period 1952–56 before moving to Harvard University and then, in 1958, to Cornell University, where he spent the rest of his career. First he was head of the Center for Radiophysics and Space Research there, and then professor of astronomy. Soon after, he became involved in the Apollo space programme (but never tired of pointing out that an unmanned effort could yield much more science for the same money). He designed the astronauts' camera, and also raised the possible problem of the Moon's surface.

The darkness of the Moon was a puzzle. Gold suggested that the Moon is covered by a layer of dust resulting from the ceaseless bombardment of its surface by Solar System debris. But his analysis did not determine the dust's thickness. He warned that a very thick layer could act like quicksand. This suggestion made NASA send a robotic device to analyse the Moon's surface conditions. Although Gold's prediction of a dust layer was fully confirmed, he received little credit for it. Only his warning of the dangers of a deep layer was remembered.

In the late 1960s, radioastronomers discovered intense flashes of radio waves recurring at precise intervals and arriving from specific directions. Their sources were puzzling. Gold proposed the now universally accepted solution: that a very dense object (a neutron star), spinning about its axis, emits radiation along a line fixed in the object. The beam of radiation from this 'pulsar' hits us once in each rotation.

Gold's last major intervention was in geology. He thought that large amounts of hydrocarbons (especially methane) had been part of the Earth as it formed, and that the Earth has been 'outgassing' ever since. The methane rises through a layer of living organisms that he called 'the deep hot biosphere'. This corresponds to the observation of organisms living near mid-ocean ridges at high temperatures and existing on chemical energy. The rising methane acquires its organic 'baggage' there, but its origin is not organic. Most Earth scientists take issue with Gold's view. But if it is correct, reservoirs of hydrocarbons exist deep underground almost everywhere, with potentially enormous geopolitical consequences.

Tommy Gold will long be remembered as a singular scientist who stepped into any field where he thought an option was being overlooked. He was also unusual in working mainly theoretically, but using little mathematics, relying instead on his profound intuitive understanding of physics.

Hermann Bondi

Hermann Bondi is at Churchill College, Cambridge CB3 0DS, UK.

Chemistry

Green fuels pick up the pace

Green Chem. **6**, doi:10.1039/b404883k (2004)

Biodiesel is an alternative fuel derived from fatty acids in vegetable oils or animal fats. Because it contains almost no sulphur, it burns more cleanly than conventional diesel. However, the chemical reaction that breaks down the fatty acids into fuel relies on an aqueous alkali catalyst such as sodium hydroxide, making the product difficult to separate from the reaction mixture.

Robert S. Watkins *et al.* now unveil a solid catalyst for this reaction that works just as well as a conventional catalyst, yet allows much easier separation of products. Their experiments used starting materials that contain shorter carbon chains than those found in vegetable oils, as a simpler model system for screening potential catalysts.

The authors achieved a 100% yield of the product using a calcium oxide catalyst loaded with 1.23% (by weight) lithium nitrate. The catalyst could be easily filtered from the reaction mixture, and recycled with minimal loss of activity. This could open up the possibility of synthesizing biodiesel as a continuous process, rather than in batches.

Mark Peplow

Plant biology

Stretching a point

Plant Physiol. doi:10.1104/pp.104.041483 (2004)

The growth of a pollen tube from a pollen grain depends on fluxes of metal ions. In particular, a calcium gradient, with highest concentrations in the tip, is needed to sustain tube elongation. Rajiv Dutta and Kenneth R. Robinson have studied events in the Easter lily, *Lilium longiflorum*, and have identified an ion channel that is responsible for calcium influx into the growing tube, and which is activated by stretching of the cell membrane.

The lily pollen tube grows in cycles of elongation every 40 to 60 seconds, accompanied by oscillations in the concentrations of various ions. Using electrophysiological measurements, Dutta and Robinson detected a spontaneously active potassium channel in the membrane of pollen grains, as well as potassium and calcium channels that were activated when the membrane was stretched by the measuring pipette. Only stretch-activated calcium channels were detected in pollen-tube tips. Blockade of this channel inhibited tube growth.

The increase in intracellular calcium occurs slightly after the elongation phase of pollen-tube growth, in line with the idea that a stretch-activated channel is involved.

This process would also position calcium influx at the point of maximum elongation, providing a feedback loop that could account for the growth oscillations seen in lilies. It remains to be seen how potassium enters the tip.

Christopher Surridge

Earthquakes

Shock reconstruction

J. Geophys. Res. B **109**, doi:10.1029/2003JB002523 (2004)

On 29 June 1170, a powerful earthquake shook the Middle East, developing along a 1,000-km fault that runs from southern Turkey to the Red Sea. Arabic documents of the time record the impact of the quake, but the sources are relatively few and they differ in their assignment of the earthquake's epicentre.

Emanuela Guidoboni and colleagues have delved into contemporary accounts in



other languages to build a more complete picture of the event. The timing of the quake between the second and third Crusades (1147–49 and 1189–92) meant that there were people of many nationalities in the region. Among the accounts in Latin are those of William of Tyre (pictured), who wrote of the city of Tripoli being reduced to an “untidy pile of stones”; and of Amalric the First, King of Jerusalem, who informed the King of France by letter that “nearly all the castles and cities situated between Tripoli and Antioch” were destroyed. Michael the Syrian — writing in Syriac, an ancient dialect of Aramaic — records that the northern city of Aleppo was worst hit.

Guidoboni *et al.* used the new data to reconstruct the probable epicentre of the earthquake, in western Syria. They estimate that the quake's magnitude was 7.7, assuming it was a single event: the account by William of Tyre suggests that the region might have been hit twice.

Alison Wright

Diagnostics

Cool amplification of DNA

EMBO Rep. doi:10.1038/sj.embor.7400200 (2004)

Many diagnostic tests depend on the amplification of a particular DNA sequence for further assays, the main method used being the polymerase chain reaction (PCR). Because the enzyme that copies DNA needs a single-stranded template to make the second strand, the duplex DNA that is generated must be separated by a brief high-temperature pulse in every cycle of amplification. A drawback of this process is that the thermal cycling equipment is both expensive and energy-demanding.

Myriam Vincent *et al.* describe a method for DNA amplification, called helicase-dependent amplification, that doesn't require high temperatures. In their protocol, duplex DNA is separated into single strands by a helicase — a protein that travels along DNA, breaking the bonds between strands. By adding a helicase to a PCR-like reaction, the amplification process can be conducted at constant, lower temperature.

Like PCR, Vincent and colleagues' method can amplify DNA sequences a million-fold. Moreover, the authors suggest how the technique may be improved, making helicase-dependent amplification a potential alternative to standard PCR, particularly where use of a thermocycler is not possible.

Angela K. Eggleston

Nanotechnology

Nanotube diodes

Appl. Phys. Lett. **85**, 145–147 (2004)

Diodes are fundamental components of modern electronics. In the most common type of diode, two different kinds of semiconductor material are joined together — one with an excess of electrons and the other with an excess of positively charged ‘holes’ for the electrons to fall into. This ensures that current can only flow in one direction through the device.

J. U. Lee *et al.* have made diodes from single-walled carbon nanotubes that measure 0.5–3 nm across and operate efficiently at room temperature. Most semiconductors need a smattering of extra chemical elements to change the number of electrons or holes inside them, but with nanotubes the same effect can be achieved through ‘electrostatic doping’ — that is, using an external electric field to change the properties of two different parts of the same nanotube. Even better, this means that the field can easily be reversed, changing the preferred direction of current flow.

Lee *et al.* suggest that their nanotube diodes could also function as light-emitting diodes at infrared wavelengths, and could therefore be useful in the development of nanoscale optoelectronic devices.

Mark Peplow

Coalition among male fiddler crabs

Seeing off a neighbour's intruder may be easier than negotiating with a larger usurper.

Until now, no compelling evidence has emerged from studies of animal territoriality to indicate that a resident will strategically help a neighbour to defend its territory against an intruder^{1,2}. We show here that territory-owning Australian fiddler crabs will judiciously assist other crabs in defending their neighbouring territories. This cooperation supports the prediction³ that it is sometimes less costly to assist a familiar neighbour than to renegotiate boundaries with a new, and possibly stronger, neighbour⁴.

In Darwin, Australia, the sexually dimorphic fiddler crab *Uca mjoebergi* lives in mixed-sex colonies on intertidal mudflats. Males have one greatly enlarged claw (Fig. 1), which is used both for attracting females and as a weapon^{5,6}. Each crab defends an all-purpose territory containing a central burrow that is used as a refuge during high tide. Males mainly fight other males and vigorously defend their territory against wandering 'floaters' that are seeking a new burrow. They also repel neighbours that encroach on their territory.

Fights involving three males have been recorded, but the territorial status of the participants was unknown⁵. We therefore tracked 268 floaters until we saw them fighting a territorial male. We recorded 17 cases in which a resident that was fighting an intruding floater was joined by an immediate neighbour (henceforth called an 'ally'). These fights always occurred at the resident's burrow entrance so the ally had to vacate his own territory temporarily to join the fight. The two neighbours never fought each other. They pushed or grappled only with the intruder.

Allies helped when they were most likely to have a beneficial effect. In unassisted fights, the smaller the resident was compared with the floater, the greater was the likelihood that he would be evicted (logistic regression, $\chi^2_1 = 36.4$, $P < 0.001$, $n = 251$). In 94% of cases when assistance was provided, the resident was smaller than the intruding floater, but this was only true in 51% of cases when no assistance was given (Fisher's exact test, $P = 0.001$, 16/17 compared with 128/251). Males therefore provided assistance when their neighbour was more likely to lose his territory.

In addition, the ally was generally larger than the floater (binomial test, $P = 0.013$, 14/17; mean $\pm$ s.d. for claw length, 21.0 ± 2.9 mm compared with 18.5 ± 4.0 mm) and therefore more likely than his smaller



Figure 1 The Australian fiddler crab uses its enlarged claw to defend non-threatening neighbours from stronger, encroaching males.

neighbour (14.8 ± 4.0 mm) to defeat the floater. The ally was always larger than the assisted neighbour (binomial test, $P < 0.0001$, 17/17). The net result was that assisted residents were evicted during the contests significantly less often than unassisted residents (logistic regression, $\chi^2_1 = 9.6$, $P = 0.002$, $n = 268$, 12% compared with 29%).

Helping is potentially costly. Allies leave their territory, which increases the risk that their own burrow will be usurped⁷, and fighting is energetically expensive and can result in claw loss^{5,6}. So why do territory owners help neighbours? First, we can exclude direct reciprocal altruism⁸ because allies were always larger than the males they assisted. Second, there is no evidence that floaters pose a delayed threat to allies. When we continued to track a floater after his first fight, we never saw him fight his previous opponent's neighbour; floaters moved 73.7 ± 34.9 cm between successive fights, compared with a mean distance between neighbours of ± 10 cm ($n = 30$ floaters).

It has been predicted that residents will form territorial coalitions when losing neighbours is costly⁹. To test this, we located 20 large males with a smaller neighbour (size-matched to observed helper-resident pairs) and watched them for 15 minutes. In no case did we observe threat displays or fights between neighbours. To mimic neighbour eviction or retention, we then either replaced the smaller neighbour with an intermediate-sized male ($n = 10$), or caught the smaller

neighbour and then returned him to his territory ($n = 10$). Once both territory owners were present on the surface, we watched them for another 15 minutes. We observed no threat displays or fights when the smaller neighbour stayed. After replacement, however, the larger male approached and fought the unfamiliar neighbour in nine out of ten cases (Fisher's exact test, $P < 0.001$).

Territorial coalitions in *U. mjoebergi* seem to be due to by-product mutualism⁹: the ally pays to retain an established neighbour and the neighbour keeps his territory. The circumstances under which assistance was provided appeared to involve judicious decision-making. That this occurs in an invertebrate, but has still not been reported in birds or mammals, suggests that territorial coalitions depend more on appropriate circumstances than on advanced cognitive skills.

**Patricia R. Y. Backwell,
Michael. D. Jennions**

School of Botany and Zoology, The Australian National University, Canberra ACT 0200, Australia
e-mail: pat.backwell@anu.edu.au

1. Morrell, L. J. & Kokko, H. *Behav. Ecol. Sociobiol.* **54**, 385–395 (2003).
2. Elfrström, S. T. *Anim. Behav.* **54**, 535–542 (1997).
3. Getty, T. *Am. Zool.* **27**, 327–336 (1987).
4. Stamps, J. A. & Krishnan, V. V. *Am. Nat.* **157**, 154–169 (2001).
5. Crane, J. A. *Fiddler Crabs of the World* (Princeton Univ. Press, Princeton, New Jersey, 1976).
6. Backwell, P. R. Y., Christy, J. H., Telford, S. R., Jennions, M. D. & Passmore, N. I. *Proc. R. Soc. Lond. B* **267**, 719–724 (2000).
7. Hemmi, J. M. & Zeil, J. *Nature* **421**, 160–163 (2003).
8. Trivers, R. L. *Q. Rev. Biol.* **46**, 35–57 (1971).
9. Dugatkin, L. A. *Cooperation Among Animals* (Oxford Univ. Press, New York, 1997).

Competing financial interests: declared none.

Corrigendum

Producing decaffeinated coffee plants

Shinjiro Ogita, Hirtaka Uefuji, Yube Yamaguchi,
Nozomu Koizumi, Hiroshi Sano
Nature **423**, 823 (2003).

It has been drawn to *Nature's* attention that S. O., N. K. and H. S. are named as the inventors on a patent relevant to this work (WO2004006658, filed in July 2002), which should therefore have been declared as a competing financial interest.

brief communications arising online

► www.nature.com/bca

Evolutionary genetics: Ambiguous role of CCR5 in *Y. pestis* infection

S. J. Elvin, E. D. Williamson, J. C. Scott, J. N. Smith,
G. Pérez de Lema, S. Chilla, P. Clapham, K. Pfeffer,
D. Schlöndorff & B. Luckow (doi:10.1038/nature02822)

Evolutionary genetics

Ambiguous role of CCR5 in *Y. pestis* infectionArising from: J. Mecsas *et al.* *Nature* **427**, 606 (2004)

Mecsas and colleagues suggest that a deficiency in the chemokine receptor CCR5 in humans is unlikely to confer protection against plague, based on their study of *Yersinia pestis* infection in Ccr5-deficient mice¹. They were testing the hypothesis that a mutation in the CCR5 gene, frequently found in Caucasians, may have been selected for in the past because it provided protection against (bubonic) plague²⁻⁷; the mutation, called CCR5Δ32, is characterized by a 32-base-pair deletion. We have also tested this hypothesis by using *Y. pestis* infection in mice and, in addition, we have done phagocytosis experiments with macrophages from wild-type and Ccr5-deficient mice. Although, like Mecsas *et al.*, we did not see any difference in the survival of the two groups of mice, we did find that there was a significantly reduced uptake of *Y. pestis* by Ccr5-deficient macrophages *in vitro*. Our results indicate that the role of Ccr5 in *Y. pestis* infection may therefore be more complex than previously thought.

In humans, macrophages are targeted by *Y. pestis*, the causative agent of plague, and are therefore important for successful infection. We tested whether Ccr5 affects the uptake of *Y. pestis* by macrophages *in vitro* by using peritoneal macrophages from Ccr5^{+/+} and Ccr5^{-/-} mice in phagocytosis assays. The uptake by Ccr5^{-/-} macrophages was about 30-fold lower than that by Ccr5^{+/+} macrophages (Fig. 1a; six independent

experiments). Our preliminary results indicate that the uptake of *Yersinia pseudotuberculosis* by macrophages from Ccr5^{-/-} mice is much less inhibited in similar experiments (Fig. 1b), suggesting that the inhibition may be specific to *Y. pestis*.

To test the effect of Ccr5 on survival after *Y. pestis* infection, groups of specific pathogen-free Ccr5^{+/+} and Ccr5^{-/-} mice were challenged with lethal inocula of *Y. pestis* GB, a highly virulent strain isolated from a fatal human case of plague. However, there was no significant difference in survival between the groups, even after infection with a low dose of two colony-forming units (CFU) (Fig. 1c).

Our survival data are in agreement with those of Mecsas *et al.*¹, although we used a strain of *Y. pestis* with a different degree of virulence (GB rather than KIM), mice with a different genetic background (C57BL/6 rather than BALB/c) and a different route of infection (subcutaneous rather than intravenous). Our results show that Ccr5^{-/-} mice are not protected against infection with a fatal human isolate of *Y. pestis* and succumb at the same rate as Ccr5^{+/+} mice.

Although these results seem to disprove the 'plague hypothesis', some doubts remain. We consistently observed a marked reduction in the uptake of *Y. pestis* by Ccr5^{-/-} macrophages *in vitro* that appears to be specific to this species of *Yersinia*. The *Y. pestis* strain that caused the great plague pandemic in the fourteenth century was probably quite

different from the twentieth-century isolate used for the infection experiments discussed here. Genome analysis indicates that *Y. pestis* evolved rapidly from an enteric organism, which was spread by the faecal–oral route, to a flea-transmitted pathogen of rodents and humans, with acquisition of novel virulence mechanisms along the way^{8,9}.

In addition, the pathogenesis of *Y. pestis* infection may not be comparable when delivered by injection of mice in the laboratory rather than by flea-borne transmission to humans¹⁰, because infection may be more rapid and acute. The dose of plague bacteria delivered by flea-borne transmission is likely to be more variable and the outcome of infection to depend on an interaction between the pathogen, vector and mammalian host. A previous infection leading to preactivation of the host's immune system would change the course of a subsequent *Y. pestis* infection — as would be expected in people living in the Middle Ages, who were constantly encountering all kinds of infection and in whom a resistance to plague could have developed in association with the CCR5Δ32 mutation.

Under these circumstances, firm conclusions cannot be drawn from the negative results obtained in Ccr5-deficient mice. Taking all these arguments into consideration, the data on the role of CCR5 in *Y. pestis* infection are still inconclusive because the situation seems to be more complex than previously anticipated.

Stephen J. Elvin*, **E. Diane Williamson***,
Joanne C. Scott*, **Jeremy N. Smith***,
Guillermo Pérez de Lema†, **Silvia Chillat†**,
Paul Clapham‡, **Klaus Pfeffer§**,
Detlef Schlöndorff††, **Bruno Luckow†**

*Defence Science and Technology Laboratories,
Porton Down, Salisbury SP4 0JQ, UK
e-mail: dewilliamson@dstl.gov.uk

†Klinikum der Universität München, Medizinische
Poliklinik–Innenstadt, 80336 München, Germany

‡University of Massachusetts, Worcester,
Massachusetts 01605, USA

§Heinrich-Heine-Universität Düsseldorf, 40225
Düsseldorf, Germany

doi:10.1038/nature02822

1. Mecsas, J. *et al.* *Nature* **427**, 606 (2004).

2. Blanpain, C., Libert, F., Vassart, G. & Parmentier, M. *Recept. Chann.* **8**, 19–31 (2002).

3. Samson, M. *et al.* *Nature* **382**, 722–725 (1996).

4. Liu, R. *et al.* *Cell* **86**, 367–377 (1996).

5. Dean, M. *et al.* *Science* **273**, 1856–1862 (1996).

6. Stephens, J. C. *et al.* *Am. J. Hum. Genet.* **62**, 1507–1515 (1998).

7. Libert, F. *et al.* *Hum. Mol. Genet.* **7**, 399–406 (1998).

8. Parkhill, J. *et al.* *Nature* **413**, 523–527 (2001).

9. Wren, B. W. *Nature Rev. Microbiol.* **1**, 55–64 (2003).

10. Jarrett, C. O., Sebbane, F., Adamovics, J. J., Andrews, G. P. & Hinnebusch, B. J. *Infect. Immun.* **72**, 2052–2056 (2004).

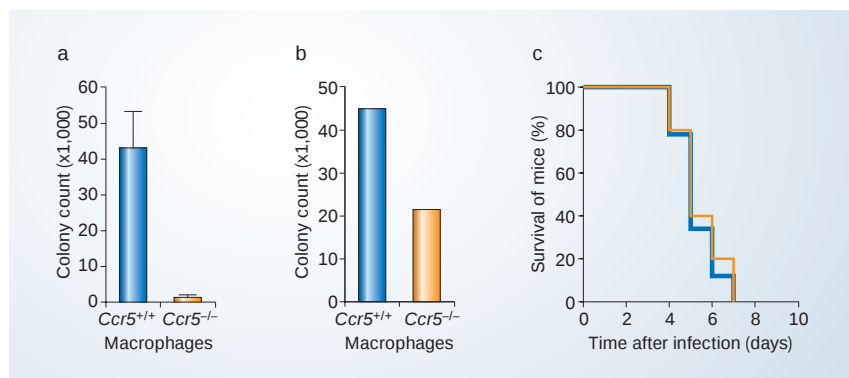


Figure 1 Ccr5 influence on the uptake of bacteria by macrophages *in vitro* and on the survival of mice infected with *Yersinia pestis*. **a, b**, The intracellular bacteria recovered from peritoneal macrophages isolated from C57BL/6 Ccr5^{+/+} and Ccr5^{-/-} mice and incubated (1×10^6 cells for 1 h at 37 °C) with **a**, *Y. pestis* GB (multiplicity of infection, 10 colony-forming units (CFU); mean $\pm$ s.e.m.) or **b**, *Y. pseudotuberculosis* strain IP32953. Gentamycin was used to kill extracellular bacteria. **c**, Survival of C57BL/6 Ccr5^{+/+} mice (blue; $n=9$) and Ccr5^{-/-} mice (orange; $n=10$) after challenge with 2 CFU *Y. pestis* GB (*Biovar orientalis*, Pgm⁺, LcrV⁺; median lethal dose is 1 CFU) subcutaneously.

Developmental plasticity and human health

Patrick Bateson¹, David Barker², Timothy Clutton-Brock³, Debal Deb⁴, Bruno D'Udine⁵, Robert A. Foley⁶, Peter Gluckman⁷, Keith Godfrey², Tom Kirkwood⁸, Marta Mirazón Lahr⁶, John McNamara⁹, Neil B. Metcalfe¹⁰, Patricia Monaghan¹⁰, Hamish G. Spencer¹¹ & Sonia E. Sultan¹²

¹Sub-Department of Animal Behaviour, University of Cambridge, High Street, Madingley, Cambridge CB3 8AA, UK

²MRC Environmental Epidemiology Unit, Southampton General Hospital, University of Southampton, SO16 6YD, UK

³Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

⁴Centre for Interdisciplinary Studies, 9 Old Calcutta Rd, Barrackpore, Kolkata 700123, India

⁵Departimento di Fisiologia Vegetale, Università di Udine, via del Cottonificio 108, 33100 Udine, Italy

⁶Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Downing Street, Cambridge CB2 3DZ, UK

⁷Liggins Institute and National Research Centre for Growth and Development, University of Auckland, Private Bag 92019, New Zealand

⁸University of Newcastle, School of Clinical Medical Sciences-Gerontology, Henry Wellcome Laboratory for Biogerontology Research, Institute for Ageing and Health, Newcastle upon Tyne NE4 6BE, UK

⁹Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK

¹⁰Division of Environmental & Evolutionary Biology, Graham Kerr Building, Glasgow University, Glasgow G12 8QQ, UK

¹¹Department of Zoology, University of Otago, PO Box 56, Dunedin, New Zealand

¹²Department of Biology, Wesleyan University, Middletown, Connecticut 06459-0170, USA

Many plants and animals are capable of developing in a variety of ways, forming characteristics that are well adapted to the environments in which they are likely to live. In adverse circumstances, for example, small size and slow metabolism can facilitate survival, whereas larger size and more rapid metabolism have advantages for reproductive success when resources are more abundant. Often these characteristics are induced in early life or are even set by cues to which their parents or grandparents were exposed. Individuals developmentally adapted to one environment may, however, be at risk when exposed to another when they are older. The biological evidence may be relevant to the understanding of human development and susceptibility to disease. As the nutritional state of many human mothers has improved around the world, the characteristics of their offspring—such as body size and metabolism—have also changed. Responsiveness to their mothers' condition before birth may generally prepare individuals so that they are best suited to the environment forecast by cues available in early life. Paradoxically, however, rapid improvements in nutrition and other environmental conditions may have damaging effects on the health of those people whose parents and grandparents lived in impoverished conditions. A fuller understanding of patterns of human plasticity in response to early nutrition and other environmental factors will have implications for the administration of public health.

Adaptive responses

Striking evidence from a number of disciplines has focused attention on the interplay between the developing organism and the circumstances in which it finds itself^{1,2,3}. Fields of research as diverse as evolutionary ecology, behavioural development, life-history theory, molecular biology and medical epidemiology have converged on the key finding that a given genotype can give rise to different phenotypes, depending on environmental conditions⁴. Many organisms can express specific adaptive responses to their environments. Such responses can include immediate, short-term changes in physiology and behaviour. However, responses to the environment may be expressed in the offspring rather than in the parent.

The freshwater crustacean *Daphnia* yields a classic example: offspring whose mother has been exposed to the chemical traces of a predator are born with a defensive 'helmet' that protects them against the predator. This structure can, however, be a liability in a predator-free environment, where its construction cost reduces competitive success relative to non-helmeted individuals⁵. Such phenotypic mismatches between the offspring's phenotype and its current environment can be costly in terms of both survival and reproductive success. The desert locust (*Schistocerca gregaria*) provides another well-known example of developmental plasticity. Under low density conditions, the locust is cryptic, shy, nocturnal and sedentary. Under crowded conditions, the individuals become increasingly conspicuous, gregarious and diurnal over several generations and then migrate in enormous swarms⁶. Both of these examples demonstrate how the impact of the environment experienced by one generation can shape the development and behaviour of the next.

Not all of the effects of the environment are adaptive. Variation

within a species may be affected by temperature, acidity, nutrient and water availability, population density, the presence of pathogens or predators, and exposure to toxins. Different phenotypes may reflect inevitable physical or chemical constraints. For example, reduced metabolic rates caused by low temperature will influence growth rate and body size. Environmental events may disrupt developmental processes, leading to abnormalities. If conditions for development are not optimal, the individual may still be able to cope, but at a cost to its future reproductive success⁷. The mature phenotype is different from one expressed under optimal conditions and is not as well adapted to adult life as it could have been. Such cases are clearly distinguishable from those where the expressed phenotype is better adapted to local conditions than alternatives^{8,9}. Many organisms have evolved to express specific adaptive responses to their environments. Such responses can include short-term changes in physiology and behaviour, as well as long-term adjustments to conditions predicted by the state of the environment when the organism is in its early stages of growth. In mice (*Mus musculus*), food restriction can slow ageing by enhancing cellular maintenance and repair processes, while reducing or shutting down fertility^{10,11}.

The varied developmental pathways triggered by environmental events may be induced during sensitive, often brief, periods in development. Outside these sensitive periods an environmental influence that sets the characteristics of an individual may have little or no effect¹². The reasons for plasticity being restricted to a particular period of life may be the difficulties of reversing developmental processes or the costs in terms of survival or reproductive success of changing the characteristics of the adult organism¹³.

As in *Daphnia*, parents of other species can make adjustments that tailor the phenotypic development of their offspring to match

prevailing environmental circumstances. Female birds, for example, are able to alter many aspects of egg composition, including nutrients, hormones, antioxidants, immunoglobulins, and even embryo sex, in response to food availability, levels of sibling competition and the quality of their mates¹⁴. Such maternal effects can result in the influence of a particular environmental factor on phenotypic development persisting across a number of generations, even if the factor itself has altered^{14,15}. In a number of vertebrates, environmental factors experienced by females during early development affect the growth and breeding success of their offspring. Mammalian mothers that themselves experience poor nutrition as a fetus often produce relatively light offspring during their breeding lifespan. For example, the daughters of food-restricted hamsters (*Mesocricetus auratus*), themselves reared with food freely available, produce smaller litters and relatively fewer sons than the daughters of control females that were not food-restricted¹⁶.

If the effects of past conditions produce mismatches with current, changed conditions, developmental plasticity may have an adverse effect on survival and reproductive success. Extensive theory has been developed to explain why developmental plasticity should be advantageous to particular plants and animals and not to others. Relevant factors include the variability of the environment, the presence of reliable environmental cues to produce an adaptive response and the rate of environmental change relative to the length of the life cycle of the species in question^{8,17,18,19}.

Triggers for human development

These broad considerations from many fields of biology are relevant to understanding of some critical variation in humans^{13,20}. The human baby responds to undernutrition, placental dysfunction and other adverse influences by changing the trajectory of his or her development and slowing growth. Although the fetus was thought to be well-buffered against fluctuations in its mother's condition, a growing body of evidence suggests that the morphology and physiology of the human baby is affected by the nutritional state of the mother^{21,22}. It is possible, therefore, that human development may involve induction of particular patterns of development by cues that prepare the developing individual for the type of environment in which he or she is likely to live. Individuals may be affected adversely if the environmental prediction provided by the mother and the conditions of early infancy prove to be incorrect¹³. Indeed, people whose birth weights were towards the lower end of the normal range and who subsequently grow up in affluent environments are at increased risk of developing coronary heart disease, type 2 diabetes and hypertension^{21,23,24}. Those born as heavier babies and brought up in affluent environments enjoy a much reduced risk. The long-term influences may arise from cues acting from before conception to infancy²⁵. The ill effects of being small, which in the short term include high death rates and childhood illness, are usually treated as yet another inevitable consequence of adversity. However, a functional and evolutionary approach derived from the rest of biology suggests that the pregnant woman in poor nutritional condition may unwittingly signal to her unborn baby that it is about to enter a harsh world. If so, this 'weather forecast' from the mother's body may result in her baby being born with characteristics, such as a small body and a modified metabolism, that help it to cope with a shortage of food. When sufficiently high levels of nutrition are available after the development of a small phenotype has been initiated, marginal benefits of rapid growth may offset the costs²⁶, but they may also trigger the health problems arising in later life.

Although adaptive responses may explain some variation in human development, it would be implausible to argue that all responses to the environment should be explained in these terms. Undernutrition, stress or hypoxia may impair normal development. Babies with low birth weight have a reduced functional capacity and fewer cells²⁷. The latter may be part of a general reduction in cell

numbers or a selective trade-off in the development of tissues that are less important to the baby, such as the kidney²⁸. Reduced numbers of nephrons at birth is a life-long deficit, as all nephrons are formed during a sensitive period of development in late gestation. The resulting increased functional demand on each individual nephron, for example by increased blood flow through each nephron, may lead to acceleration of the nephron death that accompanies normal ageing, with a consequent rise in blood pressure^{29,30}.

The diversity in past and present ecological conditions of humans is also likely to introduce complexity into the relationship between developmental prediction and later health outcome. For example, some populations may have adapted genetically to conditions of nutritional stress, especially seasonal food shortages, over a long time span, while others will have been buffered from such local evolutionary effects. The sharp increase in glucose intolerance leading to type 2 diabetes might arise from genetic differences between populations^{31,32}. The possibility of a thrifty genotype well adapted to harsh conditions is not incompatible with the plastic induction of thrifty phenotypes from a pool of uniform genotypes. However, the hypothesis that differences in susceptibility to diabetes are explained by genetic differences would not readily account for the evidence from the Dutch famine of 1944–45 that glucose intolerance is induced by maternal malnutrition during the final three months of pregnancy³³.

People who had low birth weight are resistant to the effects of insulin in moving glucose out of the blood stream into the tissues. This resistance may be adaptive for a baby receiving inadequate amounts of glucose—a way of protecting the glucose supply to the brain. However, persisting into adult life, insulin resistance leads to increasing blood glucose and type 2 diabetes develops, especially in people who have become overweight. 'Thrifty' handling of sugar becomes maladaptive if undernutrition in the womb is followed by excess in later life³⁴. Conversely, individuals with large bodies may be particularly at risk in harsh environments such as prison camps or during famines¹³. Especially striking is the evidence from a famine-exposed Ethiopian population, where the incidence of rickets was nine times greater in children who had been reported as having high birth weights than in age-matched control children³⁵. No such differences were found in children with normal birth weights.

The general hypothesis is expressed in Fig. 1. The adult health and likelihood of survival of two groups of individuals with extreme phenotypes are given for a variety of nutritional environments. For individuals whose early environment has predicted a high level of

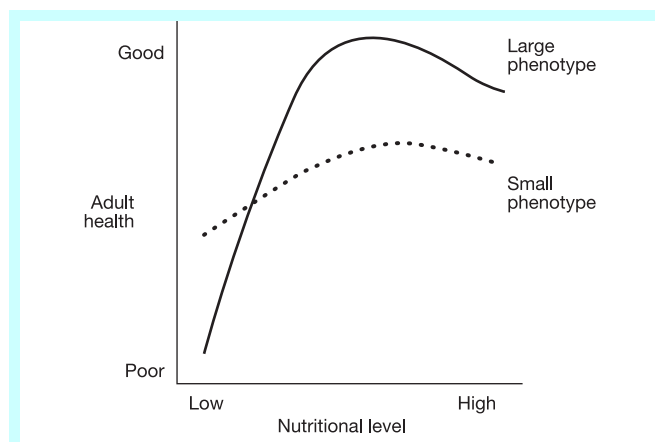


Figure 1 The hypothetical relationship between adult health and nutritional level during later development for two extreme human phenotypes that were initiated by cues received by the fetus.

nutrition in adult life and who consequently develop a large phenotype, the better the conditions, the better will be their health except perhaps at very high levels of nutrition¹⁰. For individuals whose conditions in fetal life predicted poor adult nutrition and who develop a small phenotype, the expected outcome is less clear. In very poor conditions they are expected to be healthier than those with large phenotypes. It seems plausible that their health would be benefited by some improvements in their nutritional environment in later life, but these improvements would diminish with further increases in the nutritional environment. Whether such individuals would be worse off in absolute terms than in a low nutritional environment, and the slope of the graph eventually becomes negative, is difficult to assess, but relative to the large phenotype individuals they are expected to be much less healthy.

Implications for public health

If the arguments developed here are correct, a critical public health issue is what help can be given to individuals whose characteristics were set early in life to an environment that subsequently changed. The environmental cues that trigger particular phenotypes in humans are mostly undiscovered. Nevertheless, some promising animal models have been developed, which may yield understanding of the underlying mechanisms^{36,37}. Maternal exposure to glucocorticoids in pregnancy induces hypertension and/or evidence of insulin resistance, obesity and altered muscle mass as well as alterations in the hypothalamic–pituitary–adrenal axis in the adult progeny of several experimental species. When pregnant rats are given restricted diets, their offspring are smaller at birth; but when the offspring are subsequently given plenty of food they become more obese than the offspring of mothers given an unrestricted diet^{38,39}; moreover, for mice that were malnourished *in utero*, the richer the post-natal diet, the shorter is the lifespan⁴⁰. Using animals, long-term maternal effects can be assessed in experiments that tease apart the direct genetic effects and those of the environments experienced by parents and offspring. The observed links between maternal birth weight and offspring birth weight could be due to similarity in genes, to similarity in rearing conditions imposed on both generations, or to maternal effects: poor nutrition during her own early development may reduce the capacity of a female to assimilate nutrients, and so resulting in an impaired ability to provision her offspring. Cross-fostering coupled with environmental manipulations can disentangle such effects.

The general point made here is that humans, like a great many other organisms, are capable of developing in different ways; in stable conditions, their characteristics are often well adapted to the environmental conditions in which they find themselves. An understanding of the underlying biological mechanisms (and their possible adaptive origins) should accelerate the development of appropriate interventions that promote health during childhood without compromising it later in life. □

Received 3 February; accepted 27 May 2004; doi:10.1038/nature02725.

1. Pigliucci, M. *Phenotypic Plasticity: Beyond Nature and Nurture* (Johns Hopkins Univ. Press, Baltimore, 2001).
2. Gilbert, S. F. Ecological developmental biology: Developmental biology meets the real world. *Dev. Biol.* **233**, 1–12 (2001).
3. West-Eberhard, M. J. *Developmental Plasticity and Evolution* (Oxford Univ. Press, New York, 2003).
4. Sultan, S. E. Commentary: The promise of ecological developmental biology. *J. Exp. Zool. (Mol. Dev. Evol.)* **296B**, 1–7 (2003).
5. Tollrian, R. & Dodson, S. I. in *The Ecology and Evolution of Inducible Defenses* (eds Tollrian, R. & Harvell, C. D.) 177–202 (Princeton Univ. Press, Princeton, New Jersey, 1999).
6. Pener, M. P. & Yerushalmi, Y. The physiology of locust phase polymorphism: an update. *J. Ins. Physiol.* **44**, 365–377 (1998).

7. Albon, S. D., Clutton-Brock, T. H. & Guinness, F. E. Early development and population dynamics in red deer. II. Density-independent effects and cohort variation. *J. Anim. Ecol.* **56**, 69–81 (1987).
8. Moran, N. A. The evolutionary maintenance of alternative phenotypes. *Am. Nat.* **139**, 249–278 (1992).
9. Schmitt, J., McCormac, A. C. A. H. & Smith, H. A test of the adaptive plasticity hypothesis using transgenic and mutant plants disabled in phytochrome-mediated elongation responses to neighbours. *Am. Nat.* **146**, 937–953 (1995).
10. Shanley, D. P. & Kirkwood, T. B. L. Calorie restriction and aging: a life-history analysis. *Evolution* **54**, 740–750 (2000).
11. Merry, B. J. Oxidative stress and mitochondrial function with aging—the effects of calorie restriction. *Aging Cell* **3**, 7–12 (2004).
12. Bateson, P. & Martin, P. *Design for a Life: How Behaviour Develops* (Cape, London, 1999).
13. Bateson, P. Fetal experience and good adult design. *Int. J. Epidemiol.* **26**, 561–570 (2001).
14. Lindstrom, J. Early development and fitness in birds and mammals. *Trends Ecol. Evol.* **14**, 343–348 (1999).
15. Mousseau, T. A. & Fox, C. W. *Maternal Effects as Adaptations* (Oxford Univ. Press, New York, 1998).
16. Huck, U. W., Labov, J. B. & Lisk, R. D. Food-restricting 1st generation juvenile female hamsters (*Mesocricetus auratus*) affects sex-ratio and growth of 3rd generation offspring. *Biol. Reprod.* **37**, 612–617 (1987).
17. McNamara, J. M. & Houston, A. I. State-dependent life-histories. *Nature* **380**, 215–221 (1996).
18. Tufto, J. The evolution of plasticity and nonplastic spatial and temporal adaptations in the presence of imperfect environmental cues. *Am. Nat.* **156**, 121–130 (2000).
19. Sultan, S. E. & Spencer, H. G. Metapopulation structure favors plasticity over local adaptation. *Am. Nat.* **160**, 271–283 (2002).
20. Vitzthum, V. J. A number no greater than the sum of its parts: the use and abuse of heritability. *Hum. Biol.* **75**, 539–558 (2003).
21. Barker, D. J. P. *Mothers, Babies and Health in Later Life* (Churchill Livingstone, Edinburgh, 1998).
22. Barker, D. J. P., Eriksson, J. G., Forsen, T. & Osmond, C. Fetal origins of adult disease: strength of effects and biological basis. *Int. J. Epidemiol.* **31**, 1235–1239 (2002).
23. Godfrey, K. M., Barker, D. J., Robinson, S. & Osmond, C. Maternal birthweight and diet in pregnancy in relation to the infant's thinness at birth. *Br. J. Obstet. Gynaecol.* **104**, 663–667 (1997).
24. Gluckman, P. D. & Hanson, M. A. The developmental origins of the metabolic syndrome. *Trends Endocrinol. Metab.* **15**, 183–187 (2004).
25. Eriksson, J. G. *et al.* Catch-up growth in childhood and death from coronary heart disease: longitudinal study. *Br. Med. J.* **318**, 427–431 (1999).
26. Metcalfe, N. B. & Monaghan, P. Compensation for a bad start: grow now, pay later? *Trends Ecol. Evol.* **16**, 254–260 (2001).
27. Widdowson, E. M., Crabb, D. E. & Milner, R. D. Cellular development of some human organs before birth. *Arch. Dis. Child.* **47**, 652–655 (1972).
28. Hinchliffe, S. A. *et al.* The effect of intrauterine growth-retardation on the development of renal nephrons. *Br. J. Obstet. Gynaecol.* **99**, 296–301 (1992).
29. Mackenzie, H. S. & Brenner, B. M. Fewer nephrons at birth: a missing link in the etiology of essential hypertension? *Am. J. Kidney Dis.* **26**, 91–98 (1995).
30. Keller, G., Simmer, G., Mall, G., Ritz, E. & Amann, K. Nephron number in patients with primary hypertension. *N. Engl. J. Med.* **348**, 101–108 (2003).
31. Neel, J. V. Diabetes mellitus; a thrifty genotype rendered detrimental by progress. *Am. J. Hum. Genet.* **14**, 353–362 (1962).
32. Diamond, J. The double puzzle of diabetes. *Nature* **423**, 599–602 (2003).
33. Ravelli, A. C. J. *et al.* Glucose tolerance in adults after prenatal exposure to famine. *Lancet* **351**, 173–177 (1998).
34. Hales, C. N., Desai, M. & Ozanne, S. E. The thrifty phenotype hypothesis: how does it look after 5 years? *Diabet. Med.* **14**, 189–195 (1997).
35. Chali, D., Enquessie, F. & Gesese, M. A case-control study on determinants of rickets. *Ethiop. Med. J.* **36**, 227–234 (1998).
36. Waterland, R. A. & Garza, C. Potential mechanisms of metabolic imprinting that lead to chronic disease. *Am. J. Clin. Nutr.* **69**, 179–197 (1999).
37. Chapman, C., Morgan, L. M. & Murphy, M. C. Maternal and early dietary fatty acid intake: Changes in lipid metabolism and liver enzymes in adult rats. *J. Nutr.* **130**, 146–151 (2000).
38. Jones, A. P. & Friedman, M. I. Obesity and adipocyte abnormalities in offspring of rats undernourished during pregnancy. *Science* **215**, 1518–1519 (1982).
39. Vickers, M. H., Breier, B. H., Cutfield, W. S., Hofman, P. L. & Gluckman, P. D. Fetal origins of hyperphagia, obesity, and hypertension and postnatal amplification by hypercaloric nutrition. *Am. J. Physiol. Endocrinol. Metab.* **279E**, 83–87 (2000).
40. Ozanne, S. E. & Hales, C. N. Lifespan—Catch-up growth and obesity in male mice. *Nature* **427**, 411–412 (2004).

Acknowledgements This article resulted from an interdisciplinary meeting held in Palazzo Cattaneo alla Ponzzone, Cremona, Italy in April 2003. We thank P. R. Imperiale for his encouragement and hospitality.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.B. (ppgb@cam.ac.uk).

Ancestral echinoderms from the Chengjiang deposits of China

D.-G. Shu^{1,2}, S. Conway Morris³, J. Han², Z.-F. Zhang² & J.-N. Liu²

¹School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China

²Early Life Institute and Department of Geology, Northwest University, Xi'an 710069, China

³Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK

Deuterostomes are a remarkably diverse super-phylum, including not only the chordates (to which we belong) but groups as disparate as the echinoderms and the hemichordates. The phylogeny of deuterostomes is now achieving some degree of stability, especially on account of new molecular data, but this leaves as conjectural the appearance of extinct intermediate forms that would throw light on the sequence of evolutionary events leading to the extant groups. Such data can be supplied from the fossil record, notably those deposits with exceptional soft-part preservation. Excavations near Kunming in southwestern China have revealed a variety of remarkable early deuterostomes, including the vetulicolians and yunnanozoans. Here we describe a new group, the vetulocystids. They appear to have similarities not only to the vetulicolians but also to the homalochoans, a bizarre group of primitive echinoderms whose phylogenetic position has been highly controversial.

One of the principal unsolved questions in metazoan phylogeny concerns the origin and earliest evolution of the echinoderms. Molecular data indicate echinoderms to be the sister group of hemichordates^{1,2}, but the two phyla are so disparate that attempts to envisage a common ancestor are highly conjectural. In large part this is because the echinoderm body plan has undergone radical reorganization. This includes a pervasive pentamery³, a unique water vascular system and calcitic stereom (mesodermal skeleton), loss of pharyngeal gill slits⁴, and major redeployment of developmental genes⁵. Whereas the Cambrian record of echinoderms is otherwise fairly extensive, many of the early forms have a bizarre appearance^{6–8}, and are the subject of continuing phylogenetic controversy. In part this is because of unresolved differences of opinion^{4,6,9} concerning possible similarities (for example, gill slits¹⁰) to other deuterostomes.

The Chengjiang fossil Lagerstätte^{11,12} has yielded important information on several key steps in early deuterostome evolution. These include stem-group forms, notably vetulicolians^{13–15} and yunnanozoans^{16,17}, as well as more familiar groups such as tunicates^{18,19} and vertebrates^{20,21}. The record of echinoderms is, however, much less satisfactory. A possible crinoid²² is more likely to be a lophophorate¹², whereas a supposed eocrinoid¹¹ has little similarity to other members of this group and may be arthropodan. The exclusion of stenohaline organisms, including echinoderms²³, in the Chengjiang biota is consistent with evidence for decreased seawater salinity. Here, however, we describe a new group, the vetulocystids, which we interpret as members of the Ambulacraria^{24,25}, specifically belonging to the stem group which also includes primitive echinoderms known as the homalochoans⁷. Two new taxa (*Vetulocystis catenata* gen. et sp. nov. and *Dianchicystis jianshanensis* gen. et sp. nov.) are recognized. Other material (all from the Haikou locality) is assigned to forms A and B (Fig. 3) on account of apparent differences in anatomy, but the relatively poor preservation makes it premature to use a formal taxonomy.

Total group Ambulacraria
Stem group Vetulocystida
Family Vetulocystidae fam. nov.
Vetulocystis catenata gen. et sp. nov.

Etymology. *vetus* (Latin) meaning old; the generic name is also based on its bag-like shape; *catena* (Latin) meaning chain; the specific name is a pun on missing link.

Holotype. Early Life Institute, Northwest University, Xi'an (ELI-Ech-04-001).

Referred material. ELI-Ech-04-002.

Locality. Shankou section, near Anning, about 30 km west of Kunming.

Horizon. Qiongzhusi (Chiungchussu) Formation, Yu'an-shan Member (*Eoredlichia* zone), Lower Cambrian.

Diagnosis. Bipartite body, globose theca and tail. Latter apparently with two segments, expanding posteriorly, with central strand, possibly intestine with terminal anus. Theca with three principal openings. Mouth, right, antero-dorsal, pyramidal with approximately 55 articulating platelets. Postero-dorsal opening, left, pyramidal with numerous ribs. Respiratory organ, right postero-dorsal, with folds.

Dianchicystis jianshanensis gen. et sp. nov.

Etymology. The generic name refers to Dianchi lake, near to the locality. The specific name is after Jianshan, the locality where the specimens are collected.

Holotype. Early Life Institute, Northwest University, Xi'an (ELI-Ech-04-003).

Referred material. ELI-Ech-04-004–006.

Locality. Specimens were collected from Jianshan, near Haikou, about 40 km south of Kunming.

Horizon. As for *V. catenata*.

Diagnosis. Similar to *V. catenata*, but tail tapering, with oblique striations. Anterior cone with ribs, but not arranged in platelets. Tail traversed by weakly developed furrows, possibly segmental.

All material was collected from the Lower Cambrian Qiongzhusi Formation, exposed in a wide area around Kunming, Yunnan. The best preserved material is assigned to *Vetulocystis* gen. nov. and *Dianchicystis* gen. nov., that differ most obviously with respect to the tail-like structure. The two specimens of *V. catenata* (Fig. 1) were collected from the Shankou section, near to Anning, whereas *Dianchicystis* (Fig. 2) is represented by four specimens from the Jianshan locality, near Haikou. All the specimens (including forms A and B) show typical Chengjiang preservation, and may owe their preservation to burial during storm events²³. Typically, specimens are isolated, but one enigmatic association (not illustrated) may represent the aggregation of two or more decayed individuals.

The anatomy of vetulocystids

Vetulocystids possess a common body plan that consists of an inflated theca and a shorter tail-like structure, the latter assumed to be posterior (Figs 1–3). Apart from the antero-posterior axis the remaining body orientations are more speculative because of the lack of unequivocal axial markers and the risk of circular reasoning in drawing comparisons to other groups. Tentatively, the surface carrying the two cones and respiratory organ is regarded as dorsal. Typically the theca is longer than it is wide, and whereas there is some taxonomic consistency in thecal shape the more globose examples may also reflect original flexibility and an ability to change shape. With the exception of form A (Fig. 3a, c) the integument is always strongly wrinkled (Figs 1a, d, f, g, 2a–h and 3d, g), but the absence of brittle failure argues against mineralization. In addition, despite the large-scale reticulation, the outlines of the wrinkles are irregular, of variable strength and overall have very variable and

inconsistent angular relationships. None of these features appears to be consistent with the theca being composed of calcareous plates separated by sutures. Apart from the two cones and the respiratory organ, the outer surface of the theca is smooth and shows no other evidence for mineralization. Mouldic impressions of the inner side in *Vetulocystis*, however, reveal a more irregular texture (Fig. 1f, g). In a number of specimens amorphous black/grey material can be seen within the thecal cavity, and probably represents otherwise indeterminate soft-parts.

The theca bears two prominent truncated cones, each with well developed radial ribs that converge on a central opening (Figs 1a, c, d, f, g, 2a–i, l and 3a, c–g). In *D. jianshanensis* the more anterior cone (Fig. 2a–d) consists of a series of plate-like structures (evidently unmineralized), each bearing a central rib and on either side one or two other ribs, so that in total the cone bore about 55 ribs. The plate-like structures appear to be attached to a common basal membrane

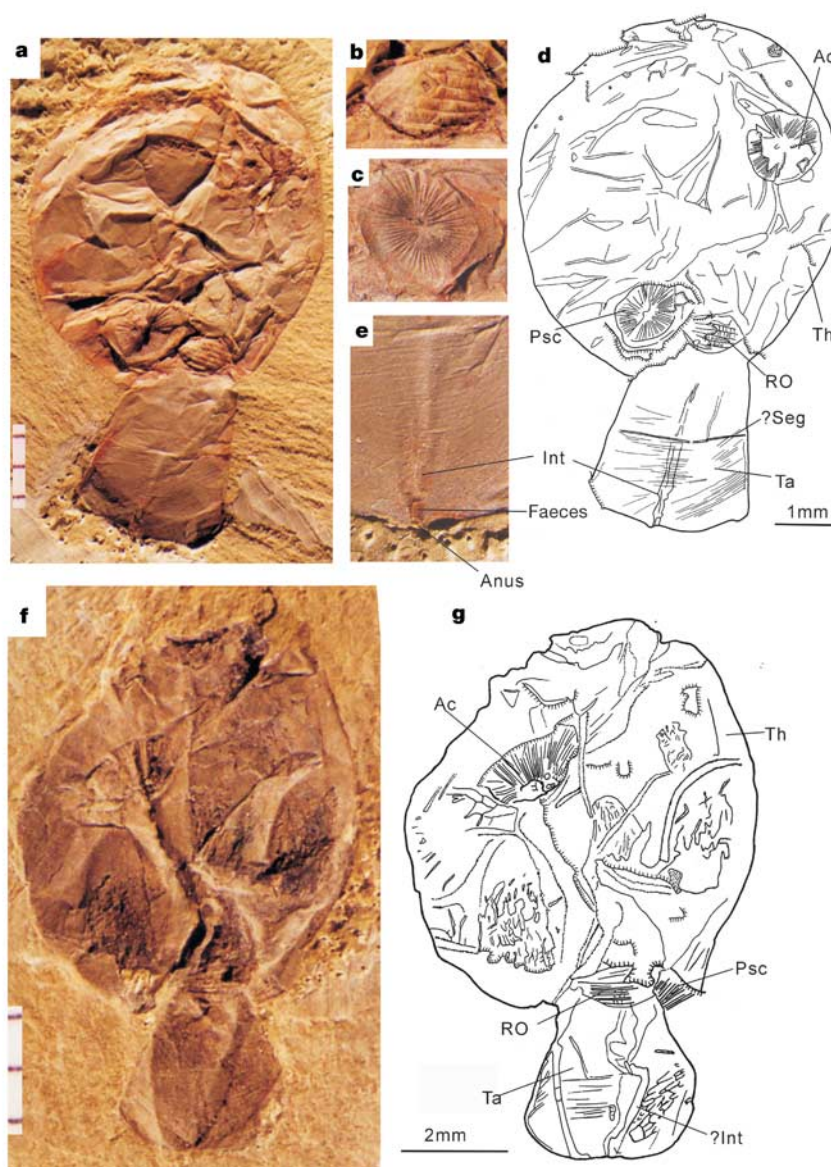


Figure 1 Two specimens of *Vetulocystis catenata* from Anning, Kunming, Yunnan. **a–e**, ELI-Ech-04-001A; **a**, entire specimen, note anterior cone is partially obscured by thecal surface; **b**, detail of respiratory organ, note cuticle missing from left-hand side; **c**, detail of the posterior cone; **d**, interpretative drawing; **e**, detail of posterior strand and sediment infill (?faeces) adjacent to ?anus. **f, g**, ELI-Ech-04-002; **f**, entire specimen,

apparently interior view of theca; **g**, interpretative drawing. Scale bars where shown on photographs are millimetric. Abbreviations: Ac, anterior cone; RO, respiratory organ; Psc, posterior cone; ?Seg, ?segment boundary; Int, possible intestine (strand); Ta, tail; Th, theca. Question marks indicate that identification of body part is tentative.

and presumably also articulated by means of connecting membranes. Towards the distal opening of the cone, finer ribs are interpolated and they presumably conferred a flexibility to assist with closure. In *V. catenata* the ribs of the anterior cone (Fig. 1a, f) appear to be individually linked to the common membrane rather than attached to plates. The anterior cone in the two named vetulocystid genera shows one apparently important difference. In *D. jianshanensis* the anterior cone is separated from the thecal surface by a basal constriction (Fig. 2a–d), whereas in *V. catenata* the equivalent structure was evidently recessed and partially obscured by the thecal surface (Fig. 1a, d). This difference may be original, but it is possible that the anterior cone was surrounded by a zone of more flexible cuticle that allowed it to be retracted into the theca. In form B (Fig. 3d, g) the anterior cone is relatively indistinct, but otherwise apparently similar to *Vetulocystis*. Adjacent to it, however, there is a striated structure. Although this may represent part of the mouth, it may alternatively be a separate organ. The equivalent cone of form B (Fig. 3a, c, e, f) differs from the other vetulocystids in being located more posteriorly.

The posterior cone in *Vetulocystis* and *Dianchicystis* is located close to the theca–tail junction (Figs 1a, c, d and 2a–h, j). It is uniformly ribbed, and presumably was moderately flexible in life. In the holotype of *D. jianshanensis* (Fig. 2a, b) this cone has a lenticular outline, but so far as can be judged in other specimens of this species and *V. catenata* it was more circular in outline. In forms A and B (Fig. 3a, c, d–g) the posterior cone is located towards the centre of the theca, and in the latter taxon is apparently composed of more massive plates (Fig. 3d, g).

The presumed respiratory organ is situated immediately adjacent to the theca–tail junction (except in form A (Fig. 3a, b, e) where it is considerably more anterior), and is transversely lenticulate (except in form B (Fig. 3d, g) where it appears to be more triangular in outline). Generally, details are difficult to discern, but in the holotype of *V. catenata* the structure is exceptionally clear (Fig. 1a, b, d). The surface appears to be cuticular and consists of a series of prominent ridges, separated by narrow grooves. In addition, each ridge has a series of transverse wrinkles. The internal structure of the respiratory organ is rather speculative. In the other specimen (Fig. 1f, g), however, the equivalent area shows a series of

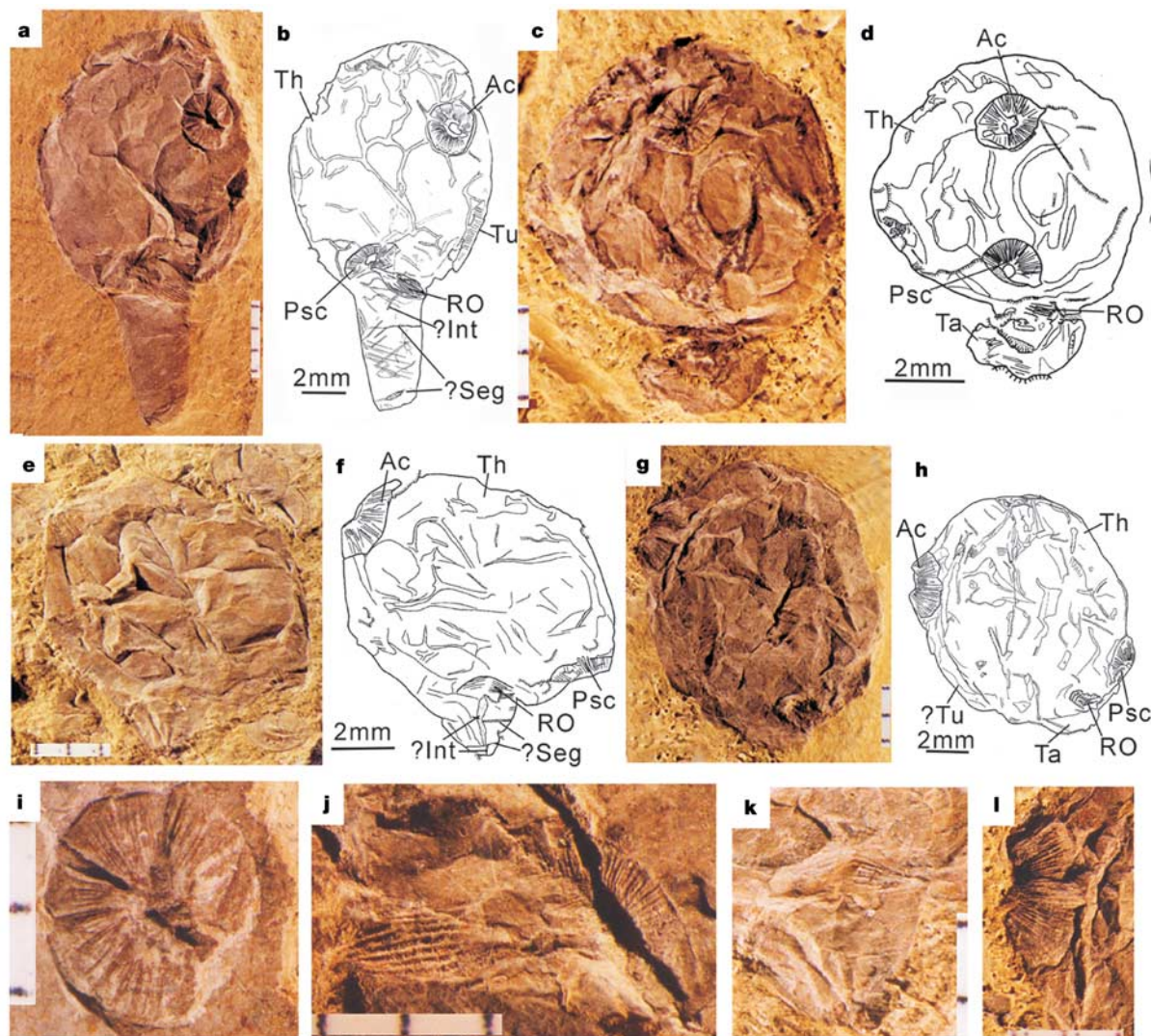


Figure 2 Four specimens of *D. jianshanensis* from Haikou, Kunming, Yunnan. **a, b, i, j**, ELI-Ech-04-003A; **a**, entire specimen; **b**, interpretative drawing; **i**, detail of anterior cone; **j**, detail of respiratory organ and posterior cone in counterpart. **c, d**, ELI-Ech-04-004A; **c**, entire specimen; **d**, interpretative drawing. **e, f, k**, ELI-Ech-04-

005; **e**, entire specimen; **f**, interpretative drawing; **k**, detail of respiratory organ and tail with possible strand. **g, h, l**, ELI-Ech-04-006; **g**, entire specimen; **h**, interpretative drawing; **l**, detail of anterior cone. Scale bars where shown on photographs are millimetric. Abbreviations: as in Fig. 1 with the addition of ?Tu, possible tube.

folded structures within a cavity. This suggests that internally the respiratory organ may have had a lamellate organization. Analysis of the functional morphology of the gill depends on various inferences and so is open to conflicting interpretations as to the possible direction of water flow. The holotype of *D. jianshanensis* (Fig. 2a, b) has a tube-like extension that extends towards, and possibly connects with, the posterior opening. In other specimens of this species (Fig. 2c–h), however, the respiratory organ and posterior opening are separated (as is even more evident in forms A and B (Fig. 3)). In *V. catenata* the respiratory organ and posterior opening are juxtaposed (Fig. 1a, d, f, g), and in this species a direct connection is conceivable.

In *V. catenata* the tail-like structure has a modest posterior expansion (Fig. 1a, d, f, g), and in the holotype has a transverse lineation. This could represent a segmental demarcation. Finer transverse striations are interpreted as cuticular folding. In the other specimen (Fig. 1f, g) a tripartite-like structure may reflect post-mortem folding, although a subdued lenticular ornamentation may be original. In the holotype a median strand extends to the posterior margin (Fig. 1a, d). It has a complex structure, but is three-dimensional and in one section appears to contain sediment (Fig. 1e). It could, therefore, represent the intestine, with gut contents, leading to a terminal anus. In the other specimen (Fig. 1f, g) the possible course of the intestine is more tentative. Identifying the strand as an intestine makes more problematic the function of the posterior cone. The tail-like structure in *D. jianshanensis* shows some important differences. This is most evident in the holotype which shows a slight taper with a blunt termination, and also two sets of oblique striations (Fig. 2a, b). In another specimen of *D. jianshanensis* the tail shows a transverse division (Fig. 2e, f), which may be segmental. The prominent

longitudinal structure is more likely to represent post-mortem folding than an intestine (Fig. 2e, f, k).

Functional biology of vetulocystids

Functional interpretations are critical in assessing both the palaeoecology and phylogenetic position of the vetulocystids. The two cones are similar (Fig. 1a, c, d, f, g and 2c, d, i, l), and identification of their respective functions is not straightforward. The anterior cone, however, is interpreted as the mouth. The closest similarity of this structure appears to lie with a number of extinct echinoderms. Thus, the putative mouth of some stylophorans^{26–28} is also pyramidal, whereas in some pentrematid blastoids the mouth is covered by a series of summit plates²⁹. An alternative location for the mouth is at the anterior end of the theca. Whereas some specimens (Fig. 2g, h) show recesses in this area, they are irregular and inconsistently preserved. The primary function of the posterior opening was presumably as the anus, possibly combined with other functions (for example, as a gonopore). It is similar to the pyramidal plated anus of many echinoderms, including the solutans^{30,31}, ctenocystoids³², cystoids³³ and some eocrinoids³⁴. The case for the lenticular structure being a respiratory organ seems to be strong. The direction of water flow and the exact site of respiratory exchange, however, are more conjectural. Here it is hypothesized that water initially entered the animal through the mouth and exited through either the respiratory organ itself or the posterior opening. Gas exchange may have been through sheets of tissue suspended in an internal cavity. Because of these uncertainties comparisons with other respiratory structures are not straightforward. There are, however, some similarities to various echinoderm respiratory structures, although these comparisons need not have specific phylogenetic implications. Moreover, the exactness of comparison depends on

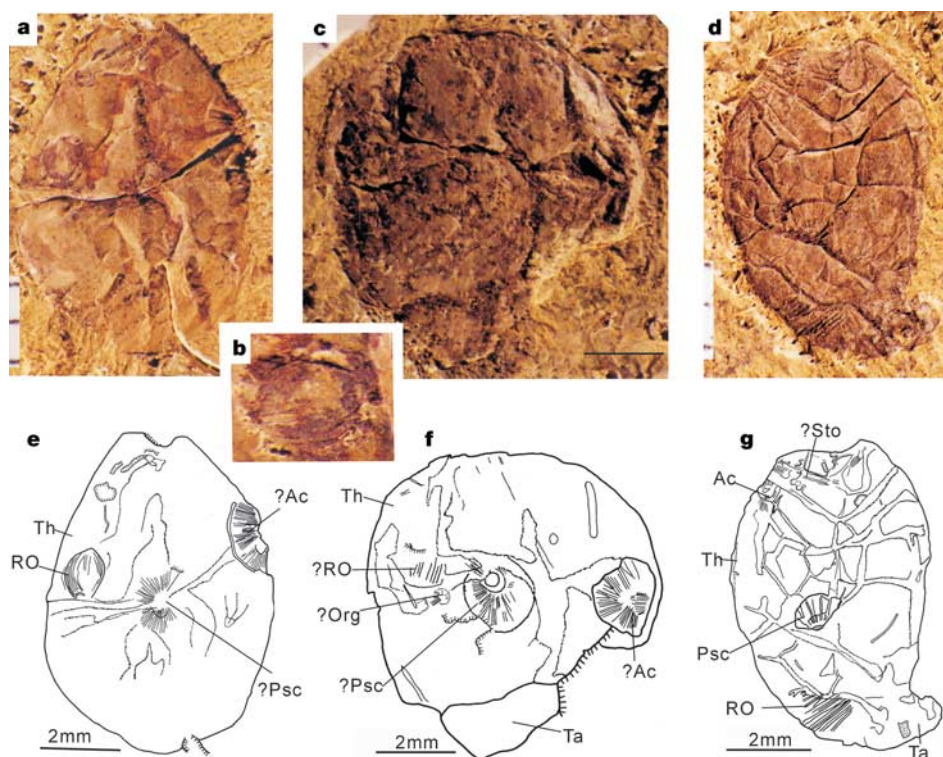


Figure 3 Form A (a–c, e, f) and form B (d, g), both from Haikou, Kunming, Yunnan. a, b, e, ELI-Ech-04-007; a, entire specimen, composite photograph of part (anterior section) and counterpart; b, detail of respiratory organ; e, interpretative drawing, composite of both part and counterpart. c, f, ELI-Ech-04-008; c, entire specimen; f, interpretative drawing. d, g, ELI-Ech-04-009A; d, entire specimen; g, interpretative drawing. Scale bars where shown on photographs are millimetric. Abbreviations: as in Fig. 1 with the additions, ?Org, unknown organ; ?Sto, striated organ.

f, interpretative drawing. d, g, ELI-Ech-04-009A; d, entire specimen; g, interpretative drawing. Scale bars where shown on photographs are millimetric. Abbreviations: as in Fig. 1 with the additions, ?Org, unknown organ; ?Sto, striated organ.

the functional arrangement of the vetulocystid respiratory organ. Thus if water exited through the organ itself, there is a similarity to the lamellar gills (lamellipores) of some cornute stylophorans^{26,27}, although in this context it is important also to note that the gill slits of the stylophorans show a wide variability of form^{26,27}. Alternatively, if gas exchange was across the outer surface of the gill with no direct access to the interior body cavity, there is an analogy to the pore rhombs of cystoids^{33,35}.

Vetulocystids were presumably semi-sessile, but a slow locomotion, possibly by sideways movement of the tail, is conceivable. Much of the time, however, they may have been stationary, possibly with the tail embedded in the sediment to provide anchorage and the theca either upright or more probably semi-prostrate with the thecal openings located on the upper surface. On the basis of the identification of the three thecal openings, it is hypothesized that suspended food particles (and possibly sea water) entered the mouth, perhaps by muscular pumping. The oral chamber may have been large and if so could have ended adjacent to the posteriorly located respiratory organ. In some specimens of *D. jianshanensis* a darker area, sometimes with faint segmentation, runs along the margin of the theca (Fig. 2a, b, g, h). It may represent a specialized organ, possibly within the anterior gut. The intestine may have looped anteriorly before exiting through the posterior cone.

The evolutionary link to echinoderms

A number of groups have adopted a body plan similar to the vetulocystids. Comparison, for example, could be made to the tunicates. Whereas this group is generally placed close to the cephalochordates and craniates, some molecular data² indicate an instability in phylogenetic placement such that a more primitive position in the deuterostomes remains conceivable. In this model, the two cones would be equivalent to the atrial and branchial siphons of tunicates. The similarity between the vetulocystid

cones and tunicate siphons is, however, only approximate, although in the extant *Chelyosoma* the siphons consist of two plated cones³⁶. A comparison could also be made with respect to the vetulocystid respiratory organ and tunicate branchial basket. The latter, however, occupies a much larger proportion of the body space, and as far as can be discerned the vetulocystid respiratory organ does not have similar microstructure. Precise comparisons between vetulocystids and tunicates, therefore, are not straightforward. In addition, tunicates from Chengjiang^{18,19} have no notable similarity to the vetulocystids.

Vetulocystids, however, have some intriguing similarities to two groups with a bipartite body plan. The first are the vetulicolians, interpreted as stem-group deuterostomes^{13,14} (see also ref. 15) possessing segmentation and gill slits. The second is an assemblage of early Palaeozoic echinoderms known as the homalozoans, which comprise the cinctans, ctenocystoids, solutes and stylophorans⁷. Similarities of the vetulocystids to the vetulicolians^{13,14} are generalized, but include a body defined by a large anterior section with a presumed pharyngeal opening and a posterior tail (Fig. 1a, d). More specific comparisons with the vetulicolians may include segmentation of the tail (Fig. 1a, d), and tentatively the identification of the strand as the intestine (Fig. 1e). The body plan of vetulocystids also has anatomical parallels to the homalozoans. Interpretations of these bizarre echinoderms are very contentious, both in terms of possible polyphyly and particularly in identification of key features that critically depend on whether the model adopted looks to echinoderm^{6,37} or chordate^{4,9,27} characteristics. Although morphologically disparate, the homalozoans arguably represent a series of key stages in early echinoderm evolution⁸, notably the retention of a bipartite body and, in cinctans and solutes³⁸, the acquisition of a feeding arm or ambulacrum (and water vascular system linked to the possibly more primitive hydropore). In addition, there is a corresponding reduction and ultimate loss of the pharyngeal gill

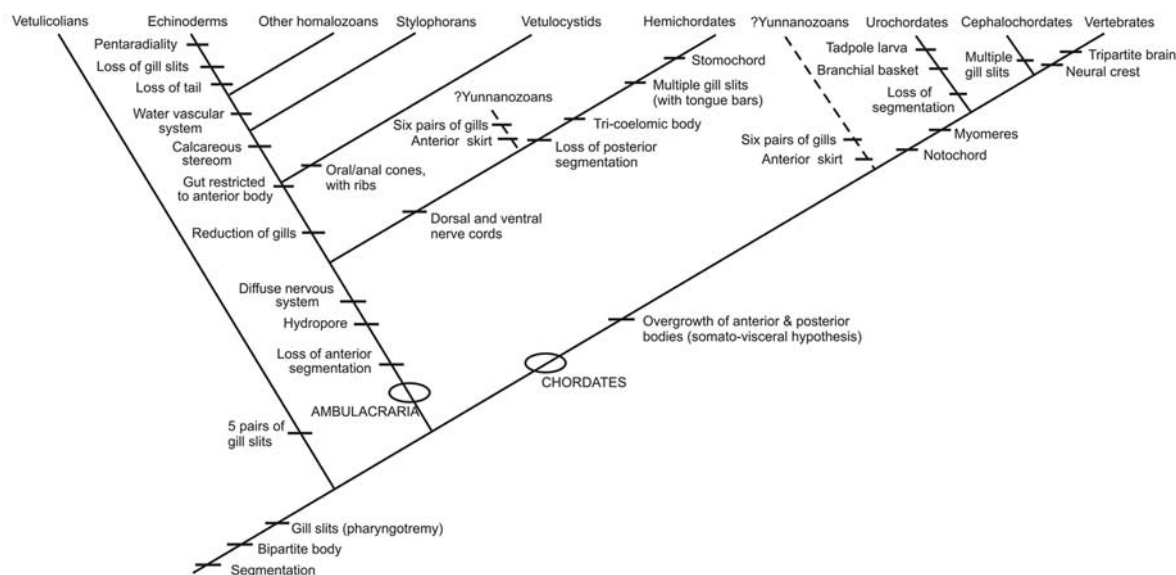


Figure 4 Phylogeny of early deuterostomes. Plesiomorphic to all deuterostomes are segmentation^{25,43} and a bipartite body, the anterior of which possesses gill slits^{24,25,44}. The posterior is a tail-like structure, segmented with an intestine and terminal anus^{13,14}. Vetulicolians^{13,14} may be the most primitive known deuterostomes, showing segmentation of the entire body, and an anterior with five pairs of gill slits. Extant members of the Ambulacraria^{24,25} are the echinoderms and hemichordates. Vetulocystids are regarded as more derived than the hemichordates, but retain the bipartite body and a respiratory organ that also characterize the most primitive echinoderms (homalozoans⁹). In more primitive Ambulacraria the gut extended along the posterior tail, but either just before or after the

bifurcation leading to the vetulocystids the gut became restricted to the anterior body. Here we depict the former possibility. All echinoderms, including homalozoans, possess a stereom, but the most primitive representatives retain gill slits. The acquisition of a water-vascular system and ambulacra was a subsequent development. The position of the extinct yunnanozoans remains controversial. Here we indicate two alternatives, either closer to the chordates¹⁷ or the hemichordates^{16,21,41,45}. In support of the latter hypothesis is the lack of evidence for key craniate features, including eyes, a complex brain, a notochord and myomeres, but the possible presence of both dorsal and ventral nerve cords.

slits. Whereas the shared possession of a bipartite body plan might be the result of evolutionary convergence, the vetulocystids possess a series of thecal openings that appear to correspond to those seen in the homalozoans. Detailed comparisons are certainly not straightforward because their relative positioning in the homalozoan groups^{8,26,38,39} is variable (as may also be the case in the vetulocystids, especially in forms A and B) presumably reflecting the widely varying thecal morphologies and inferred rearrangements of the pharyngeal cavity and coeloms. Analysis is further complicated because in only some homalozoans is there evidence for an ambulacrum. The morphology of the homalozoan tail is also variable, and presumably is connected to functional requirements that evidently shifted from being primarily locomotory^{27,31} to one of sessile attachment⁴⁰. Nevertheless, the arrangement in the vetulocystids in dorsal aspect and a clockwise direction—of mouth, respiratory organ (possibly serving as a pharyngeal opening) and anus—is similar to the arrangement in homalozoans.

A provisional phylogeny is given in Fig. 4. It is necessarily tentative because (a) the functional interpretation of vetulocystids is based on various assumptions, discussed above, (b) the position of the yunnanozoans is controversial with proposals encompassing relationships closer to the hemichordates⁴¹, pre-chordates¹⁶ and craniates¹⁷, (c) the position of the homalozoans within the echinoderms³⁷, and their inter-relationships, remain very contentious^{7,8}, and (d) the various groups ('phyla') of deuterostome are anatomically very disparate and many of the key intermediate stages are still conjectural. The scheme illustrated here, however, is rooted in the two major suppositions; the concept of the Ambulacraria^{24,25} (that is, echinoderms and hemichordates as a sister group¹) and the vetulicolians as stem-group deuterostomes^{13,14}.

In this phylogeny the vetulocystids mark a key stage in the early evolution of the Ambulacraria, specifically marking the transformation from an active vetulicolian-like organism to a semi-sessile bipartite animal which pre-figures the homalozoan echinoderms. With respect to the vetulicolians, important changes include the reduction of the respiratory organ (assumed to be equivalent to the pharyngeal opening of other deuterostomes) to a single structure adjacent to the tail, and restriction of the gut to the theca with concomitant development of oral and anal cones. On the assumption that the vetulocystids provide a key link between the vetulicolians and the homalozoans, two questions depend on presently equivocal evidence. The first revolves around the interpretation of the posterior strand as a possible intestine (Fig. 1e). If correct, this suggests that the restriction of the alimentary canal to the anterior theca took place either in vetulocystids more advanced than *V. catenata* or in the earliest homalozoans (Fig. 4). The second question concerns the evidence for posterior segmentation in some vetulocystids, and the obvious differences to the complex and variable tail structures seen in cinctans⁸, solutes^{31,39} and stylophorans^{26,27}.

In the Ambulacraria the hydropore of echinoderms is assumed to be homologous with the protocoele pore of hemichordates⁴. An equivalent structure would be difficult to recognize in the vetulocystids unless the enigmatic structures (the unknown organ and the striated organ), only seen in forms A (Fig. 3c, f) and B (Fig. 3d, g), are identified. There is, however, no evidence for a water vascular system in the vetulocystids. We suggest that it was this innovation (and the associated ambulacra), together with the evolution of a calcitic stereom skeleton, that marks the appearance of true echinoderms. It will be apparent that the discovery of the vetulocystids is further support for the 'calcichordates'²⁷ being interpreted as echinoderms^{6,42}, and not being material to the origins of the tunicates, cephalochordates and vertebrates. In contrast, however, to hypotheses that regard these curious animals as highly derived echinoderms^{6,37} it seems that they occupy a basal position⁸. □

Received 9 January; accepted 14 May 2004; doi:10.1038/nature02648.

1. Bromham, L. D. & Degnan, B. M. Hemichordate and deuterostome evolution: robust molecular phylogenetic support for a hemichordate plus echinoderm clade. *Evol. Dev.* **1**, 166–171 (1999).
2. Winchell, C. J., Sullivan, J., Cameron, C. B., Swalla, B. J. & Mallatt, J. Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. *Mol. Biol. Evol.* **19**, 762–776 (2002).
3. Sly, B. J., Hazel, J. C., Popodi, E. M. & Raff, R. A. Patterns of gene expression in the developing adult sea urchin central nervous system reveal multiple domains and deep-seated neural pentamery. *Evol. Dev.* **4**, 189–204 (2002).
4. Gee, H. *Before the Backbone: Views on the Origin of Vertebrates* (Chapman & Hall, London, 1996).
5. Wray, G. A. & Lowe, C. J. Developmental regulatory genes and echinoderm evolution. *Syst. Biol.* **49**, 28–51 (2000).
6. Lefebvre, B. Functional morphology of stylophoran echinoderms. *Palaeontology* **46**, 511–555 (2003).
7. Parsley, R. L. in *Echinoderm Research 1998* (eds Candia Carnevali, M. D. & Bonasoro, F.) 369–375 (Balkema, Rotterdam, 1999).
8. Domínguez-Alonso, P. in *Echinoderm Research 1998* (eds Candia Carnevali, M. D. & Bonasoro, F.) 263–268 (Balkema, Rotterdam, 1999).
9. Jefferies, R. P. S. A defence of the calcichordates. *Lethaia* **30**, 1–10 (1997).
10. Domínguez, P., Jacobson, A. G. & Jefferies, R. P. S. Paired gill slits in a fossil with a calcite skeleton. *Nature* **417**, 841–844 (2002).
11. Hou, X.-G., Bergström, J., Wang, H.-F., Feng, X.-H. & Chen, A.-L. *The Chengjiang Fauna: Exceptionally Well-Preserved Animals from 530 Million Years Ago* (Yunnan Science and Technology Press, Kunming, 1999).
12. Zhang, X.-L., Shu, D.-G., Li, Y. & Han, J. New sites of Chengjiang fossils: crucial windows on the Cambrian explosion. *J. Geol. Soc. Lond.* **158**, 211–218 (2001).
13. Shu, D.-G. *et al.* Primitive deuterostomes from the Chengjiang Lagerstätte (Lower Cambrian, China). *Nature* **414**, 419–424 (2001).
14. Chen, A.-L., Feng, H.-Z., Zhu, M.-Y., Ma, D.-S. & Li, M. A new vetulicolian from the early Cambrian Chengjiang fauna in Yunnan of China. *Acta Geol. Sinica* **77**, 281–287 (2003).
15. Lacalli, T. C. Vetulicolians – are they deuterostomes? *Chordates* **24**, 208–211 (2002).
16. Shu, D.-G. *et al.* A new species of yunnanozoan with implications for deuterostome evolution. *Science* **299**, 1380–1384 (2003).
17. Mallatt, J. & Chen, J.-Y. Fossil sister group of craniates: Predicted and found. *J. Morphol.* **258**, 1–31 (2003).
18. Shu, D.-G., Chen, L. & Zhang, Z.-L. An early Cambrian tunicate from China. *Nature* **411**, 472–473 (2001).
19. Chen, J.-Y. *et al.* The first tunicate from the early Cambrian of South China. *Proc. Natl Acad. Sci. USA* **100**, 8314–8318 (2003).
20. Shu, D.-G. *et al.* Head and backbone of the early Cambrian vertebrate *Haikouichthys*. *Nature* **421**, 526–529 (2003).
21. Shu, D.-G. A paleontological perspective of vertebrate origin. *Chinese Sci. Bull.* **48**, 725–735 (2003).
22. Chen, J.-Y. & Zhou, G.-Q. Biology of the Chengjiang fauna. *Bull. Natl Mus. Nat. Sci.* **10**, 11–105 (1997).
23. Babcock, L. E. & Zhang, W.-T. Stratigraphy, paleontology, and depositional setting of the Chengjiang Lagerstätte (Lower Cambrian), Yunnan, China. *Palaeoworld* **13**, 66–86 (2001).
24. Halanych, K. M. The phylogenetic position of the pterobranch hemichordates based on 18S rDNA sequence data. *Mol. Phylog. Evol.* **4**, 72–76 (1995).
25. Gee, H. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 1–14 (Taylor & Francis, London, 2001).
26. Ubahgs, G. in *Treatise on Invertebrate Paleontology Part 5, Echinodermata 1, Homalozoa-Crinozoa* (except Crinoidea) Vol. 2 (ed. Moore, R. C.) S495–S565 (Geological Society of America and Univ. Kansas, New York and Lawrence, 1967).
27. Jefferies, R. P. S. *The Ancestry of the Vertebrates* (Cambridge Univ. Press and British Natural History Museum, London, 1986).
28. Parsley, R. L. in *Echinoderms Through Time* (eds David, D., Guille, A., Féral, J.-P. & Roux, M.) 167–172 (Balkema, Rotterdam, 1994).
29. Beaver, H. H. in *Treatise on Invertebrate Paleontology Part 5, Echinodermata 1, Homalozoa-Crinozoa* (except Crinoidea) Vol. 2 (ed. Moore, R. C.) S300–S350 (Geological Society of America and Univ. Kansas, New York and Lawrence, 1967).
30. Caster, K. E. in *Treatise on Invertebrate Paleontology Part 5, Echinodermata 1, Homalozoa-Crinozoa* (except Crinoidea) Vol. 2 (ed. Moore, R. C.) S581–S627 (Geological Society of America and Univ. Kansas, New York and Lawrence, 1967).
31. Daley, P. E. J. Anatomy, locomotion and ontogeny of the solute *Castericystis vali* from the Middle Cambrian of Utah. *Geobios* **28**, 585–615 (1995).
32. Sprinkle, J. & Robison, R. A. in *Treatise on Invertebrate Paleontology Part 5, Echinodermata 2 Vol. 3* (eds Moore, R. C. & Teichert, C.) T998–T1002 (Geological Society of America and Univ. Kansas, Boulder and Lawrence, 1978).
33. Kesling, R. V. in *Treatise on Invertebrate Paleontology Part 5, Echinodermata 1, Homalozoa-Crinozoa* (except Crinoidea) Vol. 1 (ed. Moore, R. C.) S85–S267 (Geological Society of America and Univ. Kansas, New York and Lawrence, 1967).
34. Ubahgs, G. in *Treatise on Invertebrate Paleontology Part 5, Echinodermata 1, Homalozoa-Crinozoa* (except Crinoidea) Vol. 2 (ed. Moore, R. C.) S455–S495 (Geological Society of America and Univ. Kansas, New York and Lawrence, 1967).
35. Domínguez, P., Gil, D. & Torres, S. in *Echinoderm Research 1998* (eds Candia Carnevali, M. D. & Bonasoro, F.) 269–273 (Balkema, Rotterdam, 1999).
36. van Name, W. G. The North and South American ascidians. *Bull. Am. Mus. Nat. Hist.* **84**, 1–476 (1945).
37. David, B., Lefebvre, B., Mooi, R. & Parsley, R. Are homalozoans echinoderms? An answer from the extraxial-axial theory. *Paleobiology* **26**, 529–555 (2000).
38. Jefferies, R. P. S., Brown, N. A. & Daley, P. E. J. The early phylogeny of chordates and echinoderms and the origin of chordate left-right asymmetry and bilateral symmetry. *Acta Zool.* **77**, 101–122 (1996).

39. Daley, P. E. J. The anatomy of the solute *Girvanicystis batheri* (?Chordata) from the Upper Ordovician of Scotland and a new species of *Girvanicystis* from the Upper Ordovician of South Wales. *Zool. J. Linn. Soc.* **105**, 353–375 (1992).
40. Daley, P. E. J. The first solute which is attached as an adult: a mid-Cambrian fossil from Utah with echinoderm and chordate affinities. *Zool. J. Linn. Soc.* **117**, 405–440 (1996).
41. Shu, D.-G., Zhang, X. & Chen, L. Reinterpretation of *Yunnanozoon* as the earliest known hemichordate. *Nature* **380**, 428–430 (1996).
42. Jollie, M. What are the “Calcichordata”? and the larger question of the origin of chordates. *Zool. J. Linn. Soc.* **75**, 167–188 (1982).
43. Holland, L. Z., Kene, M., Williams, N. A. & Holland, N. D. Sequence and embryonic expression of the amphioxus *engrailed* gene (*AmphiEn*): the metamer pattern of transcription resembles that of its segment-polarity homolog in *Drosophila*. *Development* **124**, 1723–1732 (1997).
44. Cameron, C. B. Particle retention and flow in the pharynx of the enteropneust worm *Harrimania planktophilus*: the filter-feeding pharynx may have evolved before the chordates. *Biol. Bull.* **202**, 192–200 (2002).

45. Smith, M. P., Sansom, I. J. & Cochrane, K. D. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 67–84 (Taylor & Francis, London, 2001).

Acknowledgements Supported by the Natural Science Foundation of China, the Ministry of Science and Technology of China (D.-G.S., J.H., Z.-F.Z., J.-N.L.), The Royal Society, St John’s College, Cambridge and Cowper Reed Fund (S.C.M.). We thank M.-R. Cheng, Z.-Q. Luo, S. Last and S. Capon for technical assistance, and J.-P. Zhai, Y.-B. Ji and H.-X. Guo for help in fieldwork.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.-G. S. (elidgshu@nwnu.edu.cn).

An X-ray outburst from the rapidly accreting young star that illuminates McNeil's nebula

J. H. Kastner¹, M. Richmond¹, N. Grosso², D. A. Weintraub³, T. Simon⁴, A. Frank⁵, K. Hamaguchi⁶, H. Ozawa² & A. Henden⁷

¹Rochester Institute of Technology, Rochester, New York 14623-5604, USA

²Laboratoire d'Astrophysique de Grenoble, Université Joseph-Fourier, Grenoble, F-38041, France

³Vanderbilt University, Nashville, Tennessee 37235, USA

⁴Institute for Astronomy, Honolulu, Hawaii 96822, USA

⁵University of Rochester, Rochester, New York 14627-0171, USA

⁶Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

⁷US Naval Observatory, Flagstaff, Arizona 86002-1149, USA

Young, low-mass stars are luminous X-ray sources¹ whose powerful X-ray flares^{2–6} may exert a profound influence over the process of planet formation⁷. The origin of the X-ray emission is uncertain. Although many (or perhaps most) recently formed, low-mass stars emit X-rays as a consequence of solar-like coronal activity^{1,8,9}, it has also been suggested that X-ray emission may be a direct result of mass accretion onto the forming star^{10–12}. Here we report X-ray imaging spectroscopy observations which reveal a factor ~ 50 increase in the X-ray flux from a young star that is at present undergoing a spectacular optical/infrared outburst¹³ (this star illuminates McNeil's nebula¹⁴). The outburst seems to be due to the sudden onset of a phase of rapid accretion^{13,15,16}. The coincidence of a surge in X-ray brightness with the optical/infrared eruption demonstrates that strongly enhanced high-energy emission from young stars can occur as a consequence of high accretion rates. We suggest that such accretion-enhanced X-ray emission from erupting young stars may be short-lived, because intense star-disk magnetospheric interactions are quenched rapidly by the subsequent flood of new material onto the star.

In November 2003, V1647 Orionis, a young, low-mass star deeply embedded in the Lynds 1630 (L1630) dark cloud, which is located in the M 78 (NGC 2068) star-formation region in Orion, brightened suddenly^{13,14} in a manner that is reminiscent of optical outbursts associated with certain, less cloud-obscured, pre-main sequence (pre-MS) stars^{16–18}. Such rarely observed eruptions indicate that, for most (if not all) low-mass, pre-MS stars, long periods of relatively quiescent mass accretion at rates of 10^{-7} solar masses yr^{-1} or less are punctuated by intense, short-lived (<100 yr duration) periods during which such stars may accumulate as much as 0.01 solar masses¹⁶.

Table 1 X-ray properties of the erupting young star and nearby, field sources

No.	14 November 2002		7 March 2004		22 March 2004	
	Count rate* (counts ks^{-1})	Median E_{f} (keV)	Count rate* (counts ks^{-1})	Median E_{f} (keV)	Count rate* (counts ks^{-1})	Median E_{f} (keV)
1	8.1 ± 0.4	1.4 ± 0.1	8.7 ± 1.4	1.6 ± 0.2	5.7 ± 1.0	1.9 ± 0.3
2	32.0 ± 0.8	1.4 ± 0.1	26.2 ± 2.3	1.4 ± 0.1	25.1 ± 2.2	1.5 ± 0.1
3†	0.20 ± 0.06	2.5 ± 0.5	10.7 ± 1.4	3.6 ± 0.2	2.2 ± 0.6	2.0 ± 0.3
4‡	1.9 ± 0.2	1.7 ± 0.1	5.8 ± 1.1	1.7 ± 0.2	1.4 ± 0.5	1.1 ± 0.4

*Count rates and statistical errors within $1.5''$ radius aperture. (Errors are s.d.) 14 November 2002 data obtained in 55.9 ks effective exposure time with front-illuminated CCD ACIS-S2²²; 7 and 22 March 2004 data obtained in 5.5 ks and 4.9 ks exposure times, respectively, with back-illuminated CCD ACIS-S3. ACIS has a pixel size of 0.49 arcsec and the Chandra/ACIS combination is sensitive over the energy range 0.3–10 keV. The data were subject to standard processing by Chandra X-ray Center (CXO) pipeline software (version 7.1.1), and data reduction and analysis were performed using the CXO's CIAO package (version 3.0.2 and CALDB 2.26).

†Median energy of source photons within energy range 0.5–8 keV, and error on the median. (Errors are calculated as for s.e.m., but with median substituted for mean.)

‡Source associated with McNeil's nebula.

It is possible that pre-MS accretion processes also generate X-rays. Such high-energy radiation would occur owing to the interaction between the accretion disk and stellar magnetospheres or, perhaps, owing directly to the disk itself¹⁰. Models predict either episodic or steady reconnection of magnetic field lines occurring at the strong shear layer within which the stellar magnetosphere truncates the inner part of the accretion disk. As the fields are wound up owing to the shear they can balloon above and below the disk, eventually leading to reconnection and flares. This star-disk reconnection mechanism was proposed to explain the repeating X-ray flares of the low-mass protostar YLW 15¹⁹. Such activity may also drive well collimated outflows or jets^{20,21}.

As the first pre-MS star to be observed intensively just at the onset of an optical/infrared (IR) outburst (Fig. 1), the erupting object in L1630 affords a unique opportunity to trace the X-ray luminosity and spectral evolution of what is probably a major pre-MS accretion event. We serendipitously detected the erupting pre-MS star in X-rays in November 2002—approximately 1 year before outburst—during a deep observation of the L1630 region²² with the Advanced CCD Imaging Spectrometer (ACIS) on board the Chandra X-ray Observatory (CXO). In March 2004, following the optical outburst, we twice used CXO/ACIS to obtain short-exposure observations of a field centred on the erupting object, and detected a large increase in flux from the spatially coincident X-ray source (Table 1).

In all three CXO/ACIS observations, a point-like X-ray source is detected within $0.2''$ of the position of the erupting young star, which lies at the apex of a fan-shaped reflection nebula (McNeil's nebula; Fig. 2). The sharp increase in X-ray flux post-outburst relative to pre-outburst closely tracks that of the optical and IR brightness increase of the point-like optical/IR source (Fig. 1). Furthermore, the change in X-ray flux evidently was accompanied by a change in X-ray spectral hardness, as reflected in the mean energies of events at the three epochs of observation (Table 1). The results suggest that the source X-ray spectrum hardened considerably near the peak of the optical outburst, then softened as the outburst proceeded. The observed relationship between spectral hardness and X-ray flux indicates that the initial rise and subsequent decline in X-ray count rate is probably not due to changes in absorbing column along our line of sight but, instead,

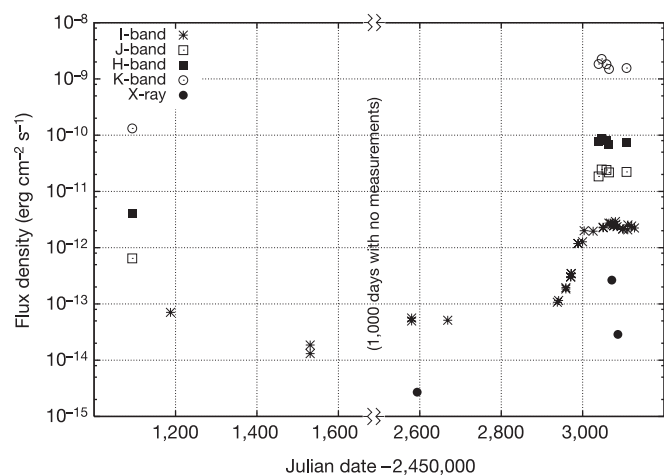


Figure 1 Near-IR and X-ray photometry of the erupting young star in L1630 obtained in the period from late 1998 through to 2004 March. CXO points are observed source fluxes calculated from the count rates in Table 1. The J- ($1.25 \mu\text{m}$), H- ($1.65 \mu\text{m}$) and K- ($2.2 \mu\text{m}$) band photometric data are from the 2MASS archive (pre-outburst) and SAAO SIRIUS, Gemini N1R1, and NOFS 1.55 m (post-outburst); I-band photometric data are from ref. 13 and B. Gary (personal communication)

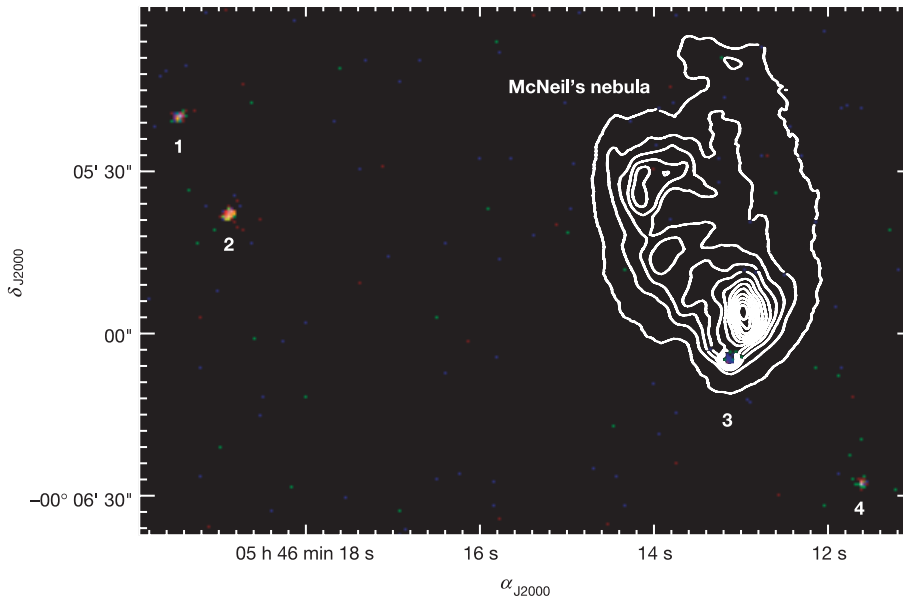


Figure 2 Chandra X-ray Observatory and visible-light images of the region surrounding McNeil's nebula. The Chandra/ACIS image (colour) is overlaid with contours from a short-exposure (10 s) R-band image of the nebula obtained with the FORS2 camera at the European Southern Observatory's Very Large Telescope. The VLT/FORS2 and CXO/ACIS images were obtained on 2004 February 18 and March 7, respectively. The Chandra image has red, green and blue colour coding for photons in the 0.5–1.5, 1.5–2.4 and 2.4–8.0 keV energy bands, respectively. Note the predominance of hard-band X-rays from the (blue) source 3 (CXO J054613.1–000604 (J2000 coordinates: right ascension, α , 05 h 46 min 13.14 s; declination, δ , $-00^{\circ} 06' 04.6''$), which lies at the apex of

McNeil's nebula). This source is spatially coincident with the IR sources 2MASS J05461313–0006048 and IRAS 05436–0007 (= LMZ 12), which have been identified with V1647 Orionis, the erupting young star now illuminating the nebula^{13,15}. Names (and J2000 coordinates) for the other sources in the field are as follows: source 1 = CXO J054619.4–000519 (05 h 46 min 19.47 s, $-00^{\circ} 05' 19.7''$; 2MASS J05461946–0005199, LkH α 301); source 2 = CXO J054618.8–000537 (05 h 46 min 18.89 s, $-00^{\circ} 05' 37.9''$; 2MASS J05461889–0005381); and source 4 = CXO J054611.6–000627 (05 h 46 min 11.61 s, $-00^{\circ} 06' 27.8''$; 2MASS J05461162–0006279).

reveals large variations in the intrinsic source X-ray luminosity and temperature. Similarly, near-IR monitoring suggests that most (~ 2.5 mag) of the abrupt ~ 3 mag change in brightness at these wavelengths can be attributed to the object's increase in intrinsic luminosity, rather than to a large change in extinction along our line of sight¹⁵.

Modelling of the X-ray spectrum obtained on 2004 March 7 indicates that the temperature of the X-ray-emitting gas (T_X) was $\sim 6 \times 10^7$ K, and that the source is viewed through an absorbing column (N_H) of $\sim 6 \times 10^{22} \text{ cm}^{-2}$ (Fig. 3). Given a gas-to-dust ratio typical of the interstellar medium, the best-fit value of N_H corresponds to dust extinction of $A_V \approx 35$ mag, whereas near-IR measurements¹⁵ yield $A_V \approx 15$ mag. This discrepancy suggests that the gas-to-dust ratio in the circumstellar environment of the young star is larger than interstellar.

Assuming no change in the absorbing column toward the X-ray source, analysis of the pre- and post-outburst X-ray observations indicates that the intrinsic source X-ray luminosity rose from $L_X \approx 3 \times 10^{29} \text{ erg s}^{-1}$ pre-outburst to $\sim 10^{31} \text{ erg s}^{-1}$ post-outburst (where we also assume the distance to L1630 is 400 pc (ref. 23) and that the pre-outburst temperature of the source was $T_X = 10^7$ K). This factor of ~ 30 increase is very similar to the increase in near-IR luminosity¹⁵, and its timing furthermore indicates a direct connection to the optical/IR event. In support of this conclusion, we note that about 15% of the $\sim 1,000$ X-ray-emitting young stars in the Orion nebula cluster (ONC) exhibited X-ray flaring (in two CXO observations spanning a combined 83 ks period²⁴); of these, only about five exhibited count rate increases of a factor of 10 or more. Thus, based solely on the most highly variable ONC X-ray sources, the probability that we have observed a large-amplitude X-ray flare that is unrelated to the optical/IR outburst is $< 3\%$. Furthermore, the erupting pre-MS star in L1630 displayed no evidence of strong

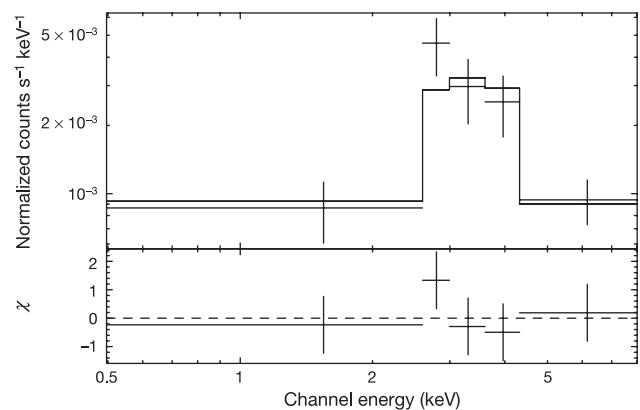


Figure 3 Chandra X-ray Observatory spectrum of the source associated with McNeil's nebula. Top: CXO/ACIS spectrum of CXO J054613.1–000604 obtained 2004 March 7 (points; errors are 1σ), with best-fit absorbed, optically thin thermal plasma model³⁰ overlaid (histogram). The best-fit model has a temperature $T_X = 5.6 \times 10^7$ K (with a lower limit of 1.5×10^7 K at 90% confidence, and unbounded upper limit) and intervening absorbing column $N_H = 5.7 \times 10^{22} \text{ cm}^{-2}$ (with a 90% confidence interval of $(3.0\text{--}10) \times 10^{22} \text{ cm}^{-2}$). Bottom: the residuals of the fit, in units of the uncertainties in the individual data points. This spectral analysis was performed with XSPEC version 11.3.

X-ray variability during the long, pre-outburst observation in November 2002.

The pre-outburst spectral energy distribution, circumstellar dust mass, and bolometric luminosity of the object in L1630 resemble those of T Tauri stars and erupting pre-MS (FU Orionis) stars²⁵, and its outburst light curve to date (Fig. 1) is very similar to those of the FU Ori prototypes at outburst onset^{16,17}. Hence, it seems reasonable

to attribute the rapid optical/IR brightness increase of this source to a sharp transition from relatively slow to exceedingly rapid accretion^{13,15,16}. The contemporaneous X-ray burst therefore is also best explained as ultimately due to accretion. The inferred post-outburst X-ray temperature was far too high for the X-rays to be generated by shocks resulting from accretion onto a low-mass, pre-MS star, however¹². Instead, the burst of X-rays was most probably generated via star-disk magnetic reconnection events that occurred in conjunction with such mass infall. This process may also launch new, collimated outflows or jets. Indeed, before its recent eruption, the pre-MS star in L1630 had been identified as the exciting source of a chain of extended emission nebulae that appears to terminate at a shock-excited Herbig-Haro object (HH 23)²⁶. The presence of these structures suggests that the present optical/IR/X-ray outburst from this object may be merely the latest of a series of such events.

If FU Ori stars are indeed among the most rapidly accreting pre-MS stars¹⁶, and X-rays from pre-MS stars can be ascribed in part to accretion, then one would naively expect all FU Ori stars to be luminous pre-MS X-ray sources. It is noteworthy, then, that before the observations reported here only two FU Ori candidates, Z CMa and L1551 IRS5, had been detected in X-rays; furthermore, both exhibit very low L_X/L_{bol} ratios^{27,28} of $\sim 10^{-6}$, compared with $L_X/L_{\text{bol}} \approx 10^{-3}$ for the erupting young star in L1630 (as estimated near the peak of its outburst). This suggests that the very large accretion rates during the steady-state phase following an FU Ori outburst eventually push the star-disk boundary sufficiently close to the stellar photosphere that the accretion becomes non-magnetospheric²⁹, thereby effectively 'quenching' X-ray emission from such objects long before the rapid accretion phase itself subsides. The precipitous drop in the X-ray flux and spectral hardness of the erupting L1630 pre-MS star, post-outburst, may signal the onset of this quenching phase, or it may indicate that the abrupt change in the nature of the star-disk interactions has triggered a phase of strong variability in both X-ray luminosity and temperature. □

Received 19 April; accepted 14 June 2004; doi:10.1038/nature02747.

- Feigelson, E. D. & Montmerle, T. High-energy processes in young stellar objects. *Annu. Rev. Astron. Astrophys.* **37**, 363–408 (1999).
- Preibisch, T., Neuhaeuser, R. & Alcalá, J. M. A giant X-ray flare on the young star P1724. *Astron. Astrophys.* **304**, L13–L16 (1995).
- Koyama, K., Hamaguchi, K., Ueno, S., Kobayashi, N. & Feigelson, E. D. Discovery of hard X-rays from a cluster of protostars. *Publ. Astron. Soc. Jpn* **48**, L87–L92 (1996).
- Grosso, N. *et al.* An X-ray superflare from an infrared protostar. *Nature* **387**, 56–58 (1997).
- Ozawa, H., Nagase, F., Ueda, Y., Dotani, T. & Ishida, M. Detection of hard X-rays from a class I protostar in the HH 24-26 region in the Orion molecular cloud. *Astrophys. J.* **523**, L81–L84 (1999).
- Imanishi, K., Koyama, K. & Tsuboi, Y. Chandra observation of the rho Ophiuchi cloud. *Astrophys. J.* **557**, 747–760 (2001).
- Feigelson, E. D., Garmire, G. P. & Pravdo, S. Magnetic flaring in the pre-main-sequence Sun and implications for the early Solar System. *Astrophys. J.* **572**, 335–349 (2002).
- Kastner, J. H. *et al.* The coronal X-ray spectrum of the multiple weak-lined T Tauri star system HD 98800. *Astrophys. J.* **605**, L49–L52 (2004).
- Stassun, K. G., Ardila, D. R., Barsony, M., Basri, G. & Mathieu, R. D. X-ray properties of pre-main-sequence stars in the Orion nebula cluster with known rotation periods. *Astron. J.* **127**, 3537–3552 (2004).
- Shu, F. H., Shang, H., Glassgold, A. E. & Lee, T. X-rays and fluctuating X-winds from protostars. *Science* **277**, 1475–1479 (1997).
- Kastner, J. H., Huenemoerder, D. P., Schulz, N. S., Canizares, C. R. & Weintraub, D. A. Evidence for accretion: High-resolution X-ray spectroscopy of the classical T Tauri star TW Hydrae. *Astrophys. J.* **567**, 434–440 (2002).
- Stelzer, B. & Schmitt, J. H. M. X-ray emission from a metal depleted accretion shock onto the classical T Tauri star TW Hya. *Astron. Astrophys.* **418**, 687–697 (2004).
- Briceno, C. *et al.* McNeil's nebula in Orion: The outburst history. *Astrophys. J.* **606**, L123–L126 (2004).
- McNeil, J. W. *IAU Circ. No.* 8284 (2004).
- Reipurth, B. & Aspin, C. IRAS 05436-0007 and the emergence of McNeil's nebula. *Astrophys. J.* **606**, L119–L122 (2004).
- Hartmann, L. & Kenyon, S. J. The FU Orionis phenomenon. *Annu. Rev. Astron. Astrophys.* **34**, 207–240 (1996).
- Herbig, G. H. Eruptive phenomena in early stellar evolution. *Astrophys. J.* **217**, 693–715 (1977).
- Herbig, G. H., Aspin, C., Gilmore, A. C., Imhoff, C. L. & Jones, A. F. The 1993–1994 activity of EX Lupi. *Publ. Astron. Soc. Pacif.* **113**, 1547–1553 (2001).
- Montmerle, T., Grosso, N., Tsuboi, Y. & Koyama, K. Rotation and X-ray emission from protostars. *Astrophys. J.* **532**, 1097–1110 (2000).
- Hayashi, M., Shibata, K. & Matsumoto, R. X-ray flares and mass outflows driven by magnetic interaction between a protostar and its surrounding disk. *Astrophys. J.* **468**, L37–L40 (1996).
- Goodson, A. P., Winglee, R. M. & Boehm, K.-H. Time-dependent accretion by magnetic young stellar

objects as a launching mechanism for stellar jets. *Astrophys. J.* **489**, 199–209 (1997).

- Simon, T., Andrews, S. M., Rayner, J. T. & Drake, S. A. X-ray and infrared observations of embedded young stars in L1630. *Astrophys. J.* (in the press); preprint at (<http://arXiv.org/astro-ph/0404260>) (2004).
- Anthony-Twarog, B. J. The H-beta distance scale for B stars—The Orion association. *Astron. J.* **87**, 1213–1222 (1982).
- Feigelson, E. D. *et al.* X-ray-emitting young stars in the Orion nebula. *Astrophys. J.* **574**, 258–292 (2002).
- Ábrahám, P. *et al.* The infrared properties of the new outburst star IRAS 05436-0007 in quiescent phase. *Astron. Astrophys.* **419**, L39–L42 (2004).
- Eisloffel, J. & Mundt, R. Parsec-scale jets from young stars. *Astron. J.* **114**, 280–287 (1997).
- Zinnecker, H. & Preibisch, Th. X-ray emission from Herbig Ae/Be stars: A ROSAT survey. *Astron. Astrophys.* **292**, 152–164 (1994).
- Favata, F., Fridlund, C. V. M., Micela, G., Sciortino, S. & Kaas, A. A. Discovery of X-ray emission from the protostellar jet L1551 IRS5 (HH 154). *Astron. Astrophys.* **386**, 204–210 (2002).
- Hartmann, L. *Accretion Processes in Star Formation* (Cambridge Univ. Press, Cambridge, 1998).
- Kaasra, J. S., Mewe, R. & Nieuwenhuijzen, H. in *UV and X-ray Spectroscopy of Astrophysical and Laboratory Plasmas* (eds Yamashita, K. & Watanabe, T.) 411 (Universal Academy Press, Tokyo, 1996).

Acknowledgements CXO observations of the erupting object in L1630 acquired in March 2004 were obtained under allocations of CXC Director's Discretionary Time. XSPEC software is maintained by NASA's High Energy Astrophysics Science Archive Research Center. The archival optical image in Fig. 2 was obtained with ESO's VLT at the Paranal Observatories under program ID 272.C-5045. We thank B. L. Gary for communicating results of I-band monitoring of the source.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.H.K. (jhk@cis.rut.edu).

Single-shot read-out of an individual electron spin in a quantum dot

J. M. Elzerman, R. Hanson, L. H. Willems van Beveren, B. Witkamp, L. M. K. Vandersypen & L. P. Kouwenhoven

Kavli Institute of Nanoscience Delft and ERATO Mesoscopic Correlation Project, Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands

Spin is a fundamental property of all elementary particles. Classically it can be viewed as a tiny magnetic moment, but a measurement of an electron spin along the direction of an external magnetic field can have only two outcomes¹: parallel or anti-parallel to the field. This discreteness reflects the quantum mechanical nature of spin. Ensembles of many spins have found diverse applications ranging from magnetic resonance imaging² to magneto-electronic devices³, while individual spins are considered as carriers for quantum information. Read-out of single spin states has been achieved using optical techniques⁴, and is within reach of magnetic resonance force microscopy⁵. However, electrical read-out of single spins^{6–13} has so far remained elusive. Here we demonstrate electrical single-shot measurement of the state of an individual electron spin in a semiconductor quantum dot¹⁴. We use spin-to-charge conversion of a single electron confined in the dot, and detect the single-electron charge using a quantum point contact; the spin measurement visibility is $\sim 65\%$. Furthermore, we observe very long single-spin energy relaxation times (up to ~ 0.85 ms at a magnetic field of 8 T), which are encouraging for the use of electron spins as carriers of quantum information.

In quantum dot devices, single electron charges are easily measured. Spin states in quantum dots, however, have only been studied by measuring the average signal from a large ensemble of electron spins^{15–20}. In contrast, the experiment presented here aims at a single-shot measurement of the spin orientation (parallel or antiparallel to the field, denoted as spin- $\uparrow$ and spin- $\downarrow$, respectively) of

a particular electron; only one copy of the electron is available, so no averaging is possible. The spin measurement relies on spin-to-charge conversion^{18,19} followed by charge measurement in a single-shot mode^{21,22}. Figure 1a schematically shows a single electron spin confined in a quantum dot (circle). A magnetic field is applied to split the spin- $\uparrow$ and spin- $\downarrow$ states by the Zeeman energy. The dot potential is then tuned such that if the electron has spin- $\downarrow$ it will leave, whereas it will stay on the dot if it has spin- $\uparrow$. The spin state has now been correlated with the charge state, and measurement of the charge on the dot will reveal the original spin state.

This concept is implemented using a structure²³ (Fig. 1b) consisting of a quantum dot in close proximity to a quantum point contact (QPC). The quantum dot is used as a box to trap a single electron, and the QPC is operated as a charge detector in order to determine whether the dot contains an electron or not. The quantum dot is formed in the two-dimensional electron gas (2DEG) of a GaAs/AlGaAs heterostructure by applying negative voltages to the metal surface gates M, R and T (Fig. 1b). This depletes the 2DEG below the gates and creates a potential minimum in the centre, that is, the dot (indicated by a dotted white circle). We tune the gate voltages such that the dot contains either zero or one electron (which we can control by the voltage applied to gate P). Furthermore, we make the tunnel barrier between gates R and T sufficiently opaque that the dot is completely isolated from the drain contact on the right. The barrier to the reservoir on the left is set²⁴ to a tunnel rate $\Gamma \approx (0.05 \text{ ms})^{-1}$. When an electron tunnels on or off the dot, it changes the electrostatic potential in its vicinity, including the region of the nearby QPC (defined by R and Q). The QPC is set

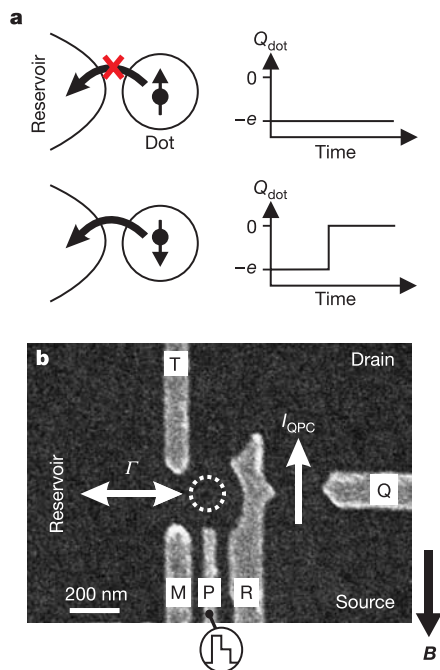


Figure 1 Spin-to-charge conversion in a quantum dot coupled to a quantum point contact. **a**, Principle of spin-to-charge conversion. The charge on the quantum dot, Q_{dot} , remains constant if the electron spin is $\uparrow$, whereas a spin- $\downarrow$ electron can escape, thereby changing Q_{dot} . **b**, Scanning electron micrograph of a device like the one used in the measurements, showing the metallic gates (T, M, P, R, Q) on the surface of a GaAs/AlGaAs heterostructure containing a two-dimensional electron gas (2DEG) 90 nm below the surface. The electron density is $2.9 \times 10^{15} \text{ m}^{-2}$. (Only the gates used in the present experiment are shown, the complete device²³ is described in Supplementary Fig. S1.) By measuring the current through the QPC channel, I_{QPC} , we can detect changes in Q_{dot} that result from electrons tunnelling between the dot and the reservoir (with a tunnel rate Γ). A magnetic field, B , is applied in the plane of the 2DEG.

in the tunnelling regime, so that the current, I_{QPC} , is very sensitive to electrostatic changes²⁵. Recording changes in I_{QPC} thus permits us to measure on a timescale of about $8 \mu\text{s}$ whether an electron resides on the dot or not (L.M.K.V. *et al.*, manuscript in preparation). In this way the QPC is used as a charge detector with a resolution much better than a single electron charge and a measurement timescale almost ten times shorter than $1/\Gamma$.

The device is placed inside a dilution refrigerator, and is subjected to a magnetic field of 10 T (unless noted otherwise) in the plane of the 2DEG. The measured Zeeman splitting in the dot¹⁹, $\Delta E_Z \approx 200 \mu\text{eV}$, is larger than the thermal energy ($25 \mu\text{eV}$) but smaller than the orbital energy level spacing (1.1 meV) and the charging energy (2.5 meV).

To test our single-spin measurement technique, we use an experimental procedure, inspired by earlier time-averaged measurements^{18,19}, that is based on three stages: (1) empty the dot, (2) inject one electron with unknown spin, and (3) measure its spin state. The different stages are controlled by voltage pulses on gate P (Fig. 2a), which shift the dot's energy levels (Fig. 2c). Before the pulse the dot is empty, as both the spin- $\uparrow$ and spin- $\downarrow$ levels are above the Fermi energy of the reservoir, E_F . Then a voltage pulse pulls both levels below E_F . It is now energetically allowed for an electron to tunnel onto the dot, which will happen after a typical time $\sim \Gamma^{-1}$. The particular electron can have spin- $\uparrow$ or spin- $\downarrow$, shown in the lower and upper diagram respectively (the tunnel rate for spin- $\uparrow$ electrons is

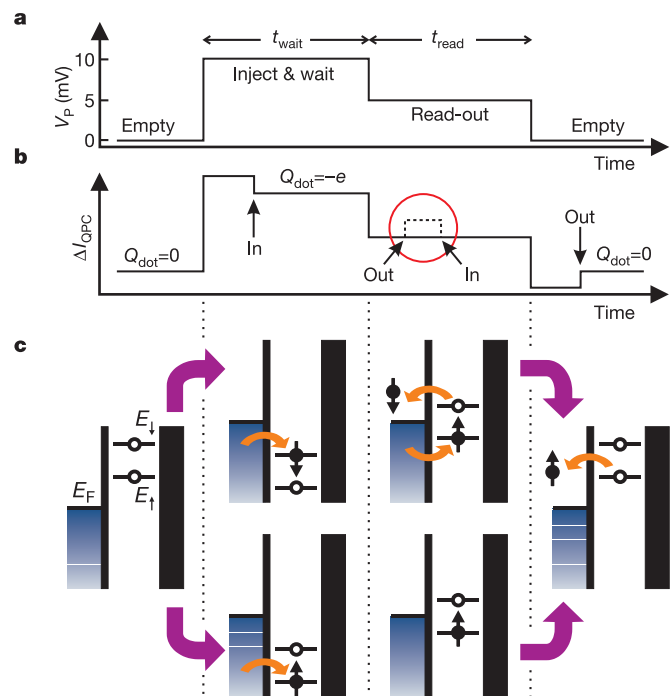


Figure 2 Two-level pulse technique used to inject a single electron and measure its spin orientation. **a**, Shape of the voltage pulse applied to gate P. The pulse level is 10 mV during t_{wait} and 5 mV during t_{read} (which is 0.5 ms for all measurements). **b**, Schematic QPC pulse-response if the injected electron has spin- $\uparrow$ (solid line) or spin- $\downarrow$ (dotted line; the difference with the solid line is only seen during the read-out stage). Arrows indicate the moment an electron tunnels into or out of the quantum dot. **c**, Schematic energy diagrams for spin- $\uparrow$ ($E_{\uparrow}$) and spin- $\downarrow$ ($E_{\downarrow}$) during the different stages of the pulse. Black vertical lines indicate the tunnel barriers. The tunnel rate between the dot and the QPC drain on the right is set to zero. The rate between the dot and the reservoir on the left is tuned to a specific value, Γ . If the spin is $\uparrow$ at the start of the read-out stage, no change in the charge on the dot occurs during t_{read} . In contrast, if the spin is $\downarrow$, the electron can escape and be replaced by a spin- $\uparrow$ electron. This charge transition is detected in the QPC current (dotted line inside red circle in **b**).

expected to be larger than that for spin- $\downarrow$ electrons²⁶, that is, $\Gamma_{\uparrow} > \Gamma_{\downarrow}$, but we do not assume this a priori.) During this stage of the pulse, lasting t_{wait} , the electron is trapped on the dot and Coulomb blockade prevents addition of a second electron. After t_{wait} the pulse is reduced, in order to position the energy levels in the read-out configuration. If the electron spin is $\uparrow$, its energy level is below E_F , so the electron remains on the dot. If the spin is $\downarrow$, its energy level is above E_F , so the electron tunnels to the reservoir after a typical time $\sim \Gamma_{\downarrow}^{-1}$. Now Coulomb blockade is lifted and an electron with spin- $\uparrow$ can tunnel onto the dot. This occurs on a timescale $\sim \Gamma_{\uparrow}^{-1}$ (with $\Gamma = \Gamma_{\uparrow} + \Gamma_{\downarrow}$). After t_{read} , the pulse ends and the dot is emptied again.

The expected QPC response, ΔI_{QPC} , to such a two-level pulse is the sum of two contributions (Fig. 2b). First, owing to a capacitive

coupling between pulse gate and QPC, ΔI_{QPC} will change proportionally to the pulse amplitude. Thus, ΔI_{QPC} versus time resembles a two-level pulse. Second, ΔI_{QPC} tracks the charge on the dot, that is, it goes up whenever an electron tunnels off the dot, and it goes down by the same amount when an electron tunnels onto the dot. Therefore, if the dot contains a spin- $\downarrow$ electron at the start of the read-out stage, ΔI_{QPC} should go up and then down again. We thus expect a characteristic step in ΔI_{QPC} during t_{read} for spin- $\downarrow$ (dotted trace inside red circle). In contrast, ΔI_{QPC} should be flat during t_{read} for a spin- $\uparrow$ electron. Measuring whether a step is present or absent during the read-out stage constitutes our spin measurement.

Figure 3a shows typical experimental traces of the pulse-response recorded after proper tuning of the d.c. gate voltages (see Supplementary Fig. S2). We emphasize that each trace involves injecting

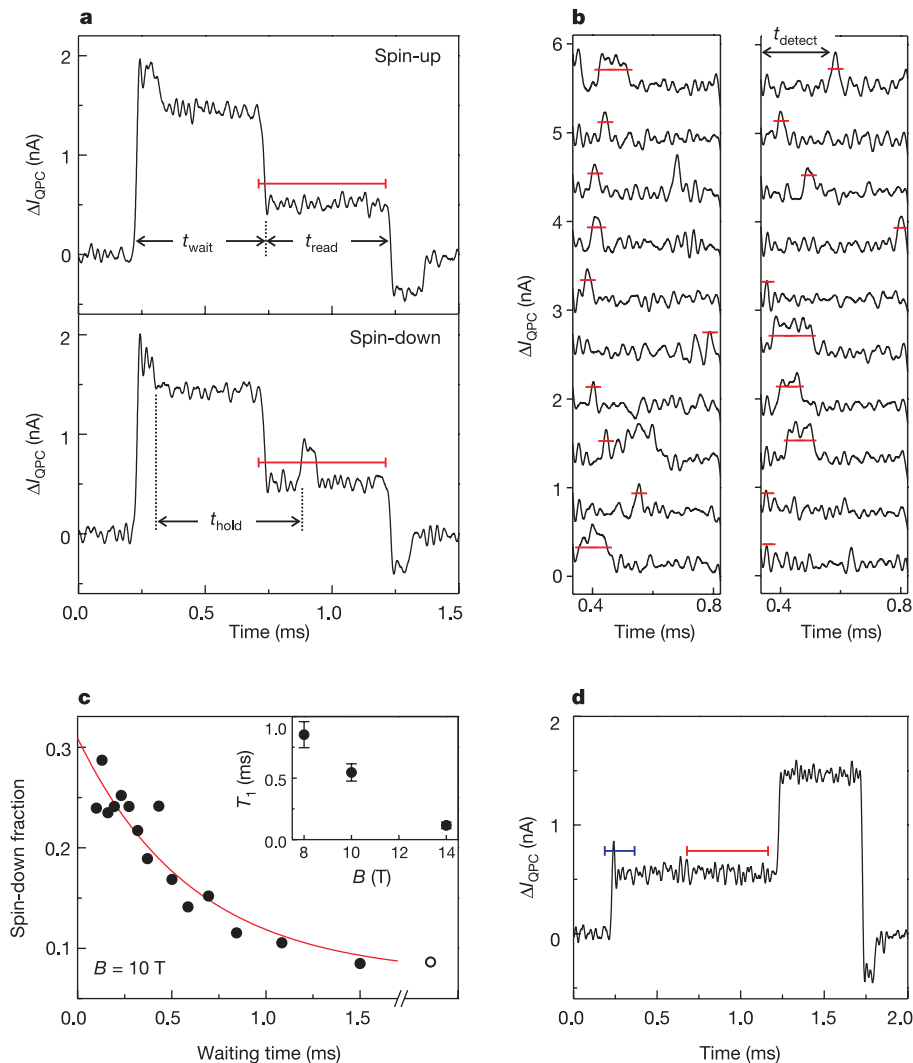


Figure 3 Single-shot read-out of one electron spin. **a**, Typical time-resolved measurements of the QPC current in response to a two-level pulse. In the top panel, an electron is injected during t_{wait} and is declared 'spin-up' during t_{read} . In the lower panel, the injected electron is declared 'spin-down' by the characteristic step which crosses the threshold (red line) during t_{read} . The total time the electron spends in the dot is defined as t_{hold} . **b**, Randomly chosen examples of traces for which the electron is declared 'spin-down' (here for the case of $t_{\text{wait}} = 0.1$ ms). Only the read-out segment is shown, and traces are offset for clarity. The actual time when ΔI_{QPC} first crosses the threshold (red line), t_{detect} , is recorded to make the histogram in Fig. 4a. **c**, Fraction of traces counted as spin-down versus waiting time, t_{wait} , out of a total of 625 traces taken for each waiting time. Rightmost point (open dot): spin-down fraction using modified pulse shape (**d**). Red

solid line: exponential fit to the data. Inset: T_1 versus B (see Supplementary Fig. S4). Error bars represent the root mean square of the standard errors obtained from exponential fits to three separate data sets. **d**, Typical QPC signal for a 'reversed' pulse, which has the same amplitudes as in Fig. 2a, but with the order of the two stages reversed. In this case injection takes place with $E_{\uparrow}$ below and $E_{\downarrow}$ above E_F (see Fig. 2c, third column), so that only a spin- $\uparrow$ electron can be injected. By recording the fraction of traces in which the current nevertheless crosses the threshold of duration t_{read} (red line), an independent measure of the 'dark count' probability is obtained (see text). This fraction is plotted as the open dot in **c**. It is used in the exponential fit with an associated value of $t_{\text{wait}} = 10$ ms (that is, much longer than the spin relaxation time). The blue threshold is used in Fig. 4b.

one particular electron onto the dot and subsequently measuring its spin state. Each trace is therefore a single-shot measurement. The traces that we obtain fall into two different classes; most traces qualitatively resemble the one in the top panel of Fig. 3a, some resemble the one in the bottom panel (and sometimes, no electron was injected during the injection stage; such cases were detected, see Supplementary Fig. S3, and ignored). These two typical traces indeed correspond to the signals expected for a spin- $\uparrow$ and a spin- $\downarrow$ electron (Fig. 2b), a strong indication that the electron in the top panel of Fig. 3a was spin- $\uparrow$ and in the bottom panel spin- $\downarrow$. The distinct signature of the two types of responses in ΔI_{QPC} permits a simple criterion for identifying the spin: if ΔI_{QPC} crosses the threshold value (red line in Fig. 3a and chosen as explained below), we declare the electron 'spin-down'; otherwise we declare it 'spin-up'. Figure 3b shows the read-out section of 20 more 'spin-down' traces, to illustrate the stochastic nature of the tunnel events.

The random injection of spin- $\uparrow$ and spin- $\downarrow$ electrons prevents us from checking the outcome of any individual measurement. Therefore, in order to further establish the correspondence between the actual spin state and the outcome of our spin measurement, we change the probability of having a spin- $\downarrow$ at the beginning of the read-out stage, and compare this with the fraction of traces in which the electron is declared 'spin-down'. As t_{wait} is increased, the time between injection and read-out, t_{hold} , will vary accordingly ($t_{\text{hold}} \approx t_{\text{wait}}$). The probability for the spin to be $\downarrow$ at the start of t_{read} will thus decay exponentially to zero, since electrons in the excited spin state will relax to the ground state ($k_B T \ll \Delta E_Z$). For a set of 15 values of t_{wait} we take 625 traces for each t_{wait} , and count the fraction of traces in which the electron is declared 'spin-down' (Fig. 3c). The fact that the expected exponential decay is clearly reflected in the data confirms the validity of the spin read-out procedure.

We extract a single-spin energy relaxation time, T_1 , from fitting the data points in Fig. 3c (and two other similar measurements) to $\alpha + C \exp(-t_{\text{wait}}/T_1)$, and obtain an average value of $T_1 \approx (0.55 \pm 0.07)$ ms at 10 T. This is an order of magnitude longer than the lower bound on T_1 established earlier¹⁹, and clearly longer than the time needed for the spin measurement (of order $1/T_1 \approx 0.11$ ms). Similar experiments at 8 T give $T_1 \approx (0.85 \pm 0.11)$ ms, and at 14 T we find $T_1 \approx (0.12 \pm 0.03)$ ms (Supplementary Fig. S4). More experiments are needed in order to test the theoretical prediction that relaxation at high magnetic fields is dominated by spin-orbit interactions^{27–29}, with smaller contributions resulting from hyperfine interactions with the nuclear spins^{27,30} (co-tunnelling is insignificant given the very small tunnel rates). For both mechanisms, T_1 is expected to decrease rapidly with magnetic field^{27–30}, in part because the energy from the spin-flip process must be absorbed by the phonon bath, which has a higher density of states at higher energies. We note that the obtained values for T_1 refer to our entire device under active operation: that is, a single spin in a quantum dot subject to continuous charge detection by a QPC.

For applications in quantum information processing it is important to know the accuracy, or fidelity, of the single-shot spin read-out. The measurement fidelity is characterized by two parameters, α and β (inset to Fig. 4a), which we now determine for the data taken at 10 T.

The parameter α corresponds to the probability that the QPC current exceeds the threshold even though the electron was actually spin- $\uparrow$, for instance due to thermally activated tunnelling or electrical noise (similar to 'dark counts' in a photon detector). The combined probability for such processes is given by the saturation value of the exponential fit in Fig. 3c, α , which depends on the value of the threshold current. We analyse the data in Fig. 3c using different thresholds, and plot α in Fig. 4b.

The parameter β corresponds to the probability that the QPC current stays below the threshold even though the electron

was actually spin- $\downarrow$ at the start of the read-out stage. Unlike α , β cannot be extracted directly from the exponential fit (note that the fit parameter $C = p(1 - \alpha - \beta)$ contains two unknowns: $p = \Gamma_{\downarrow}/(\Gamma_{\uparrow} + \Gamma_{\downarrow})$ and β). We therefore estimate β by analysing the two processes that contribute to it. First, a spin- $\downarrow$ electron can relax to spin- $\uparrow$ before the electron tunnels out. This occurs with probability $\beta_1 = 1/(1 + T_1 \Gamma_{\uparrow})$. From a histogram (Fig. 4a) of the actual detection time, t_{detect} (Fig. 3b), we find $\Gamma_{\uparrow}^{-1} \approx 0.11$ ms, yielding $\beta_1 \approx 0.17$. Second, if the spin- $\downarrow$ electron does tunnel off the dot but is replaced by a spin- $\uparrow$ electron within about 8 μ s, the resulting QPC step is too small to be detected. The probability that a step is missed, β_2 , depends on the value of the threshold. It can be determined by applying a modified ('reversed') pulse (Fig. 3d). For such a pulse, we know that in each trace an electron is injected in the

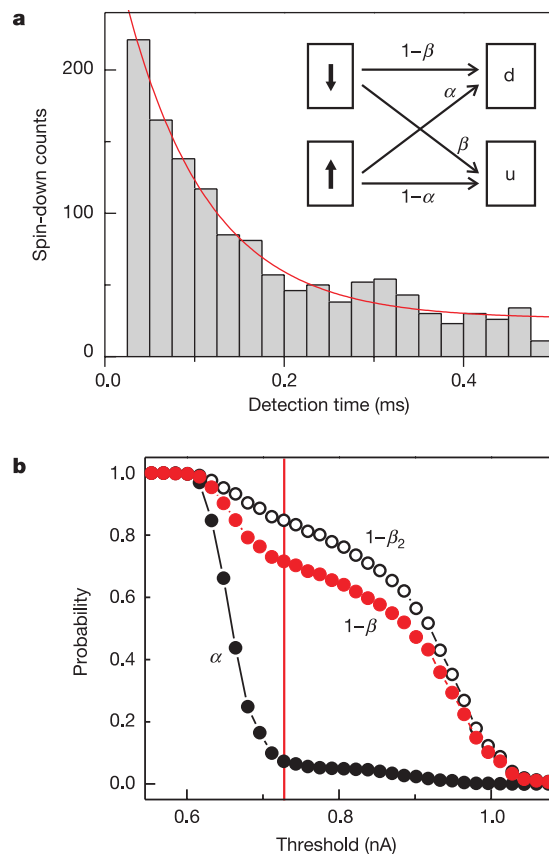


Figure 4 Measurement fidelity. **a**, Histogram showing the distribution of detection times, t_{detect} , in the read-out stage (see Fig. 3b for definition of t_{detect}). The exponential decay is due to spin- $\downarrow$ electrons tunnelling out of the dot (rate $= \Gamma_{\downarrow}$) and due to spin flips during the read-out stage (rate $= 1/T_1$). Red line: exponential fit with a decay time $(\Gamma_{\downarrow} + 1/T_1)^{-1}$ of 0.09 ms. Given that $T_1 = 0.55$ ms, this yields $\Gamma_{\downarrow}^{-1} \approx 0.11$ ms. Inset: fidelity parameters. A spin- $\downarrow$ electron is declared 'down' (d) or 'up' (u) with probability $1 - \beta$ or β , respectively. A spin- $\uparrow$ electron is declared 'up' or 'down' with probability $1 - \alpha$ or α , respectively. **b**, Closed black dots represent α , obtained from the saturation value of exponential fits as in Fig. 3c for different values of the read-out threshold. A current of 0.54 nA (0.91 nA) corresponds to the average value of ΔI_{QPC} when the dot is occupied (empty) during t_{read} . Open black dots: measured fraction of 'reverse-pulse' traces in which ΔI_{QPC} crosses the injection threshold (blue line in Fig. 3d). This fraction approximates $1 - \beta_2$, where β_2 is the probability of identifying a spin- $\downarrow$ electron as 'spin-up' owing to the finite bandwidth of the measurement set-up. Closed red dots: total fidelity for the spin- $\downarrow$ state, $1 - \beta$, calculated using $\beta_1 = 0.17$. The vertical red line indicates the threshold for which the visibility $1 - \alpha - \beta$ (separation between red and black closed dots) is maximal. This threshold value of 0.73 nA is used in the analysis of Fig. 3.

dot, so there should always be a step at the start of the pulse. The fraction of traces in which this step is nevertheless missed, that is, ΔI_{QPC} stays below the threshold (blue line in Fig. 3d), gives β_2 . We plot $1 - \beta_2$ in Fig. 4b (open dots). The resulting total fidelity for spin- $\downarrow$ is given by $1 - \beta \approx (1 - \beta_1)(1 - \beta_2) + (\alpha\beta_1)$. The last term accounts for the case when a spin- $\downarrow$ electron is flipped to spin- $\uparrow$, but there is nevertheless a step in ΔI_{QPC} due to the dark-count mechanism. In Fig. 4b we also plot the extracted value of $1 - \beta$ as a function of the threshold.

We now choose the optimal value of the threshold as the one for which the visibility $1 - \alpha - \beta$ is maximal (red line in Fig. 4b). For this setting, $\alpha \approx 0.07$, $\beta_1 \approx 0.17$ and $\beta_2 \approx 0.15$, so the measurement fidelity for the spin- $\uparrow$ and the spin- $\downarrow$ state is ~ 0.93 and ~ 0.72 , respectively. The measurement visibility in a single-shot measurement is thus at present 65%.

Significant improvements in the spin measurement visibility can be made by lowering the electron temperature (smaller α), and especially by making the charge measurement faster (smaller β). Already, the demonstration of single-shot spin read-out and the observation of T_1 of the order of 1 ms are encouraging results for the use of electron spins as quantum bits. Present experiments focus on measuring the phase coherence time, T_2 (by definition $\leq 2T_1$), by performing pulsed electron spin resonance experiments. $\square$

Received 6 April; accepted 25 May 2004; doi:10.1038/nature02693.

1. Sakurai, J. J. *Modern Quantum Mechanics* (Addison-Wesley, Reading, Massachusetts, 1994).
2. Wehrli, F. W. The origins and nature of nuclear magnetic resonance imaging. *Phys. Today* **6**, 34–42 (1992).
3. Wolf, S. A. *et al.* Spintronics: a spin-based electronics vision for the future. *Science* **294**, 1488–1495 (2001).
4. Blatt, R. & Zoller, P. Quantum jumps. *Eur. J. Phys.* **9**, 250–279 (1988).
5. Rugar, D., Budakian, R., Mamin, H. J. & Chui, B. W. Single spin detection by magnetic resonance force microscopy. *Nature* **430**, 329–332 (2004).
6. Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
7. Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* **393**, 133–137 (1998).
8. Vandersypen, L. M. K., *et al.* in *Quantum Computing and Quantum Bits in Mesoscopic Systems* (eds Leggett, A. J., Ruggiero, B. & Silvestrini, P.) 201–209 (Kluwer Academic/Plenum, New York, 2003); (<http://xxx.lanl.gov/abs/quant-ph/0207059>) (2002).
9. Xiao, M., Martin, I. & Jiang, H. W. Probing the spin state of a single electron trap by random telegraph signal. *Phys. Rev. Lett.* **91**, 078301 (2003).
10. Friesen, M., Tahan, C., Joynt, R. & Eriksson, M. A. Spin readout and initialization in a semiconductor quantum dot. *Phys. Rev. Lett.* **92**, 037901 (2004).
11. Engel, H. *et al.* Measurement efficiency and n-shot read out of spin qubits. *Phys. Rev. Lett.* (in the press); preprint at (<http://xxx.lanl.gov/abs/cond-mat/0309023>) (2003).
12. Ionićiu, R. & Popescu, A. E. Single spin measurement using spin-orbital entanglement. Preprint at (<http://xxx.lanl.gov/abs/quant-ph/0310047>) (2003).
13. Greentree, A. D., Hamilton, A. R., Hollenberg, L. C. L. & Clark, R. G. Electrical readout of a spin qubit without double occupancy. Preprint at (<http://xxx.lanl.gov/abs/cond-mat/0403449>) (2004).
14. Kouwenhoven, L. P., Austing, D. G. & Tarucha, S. Few-electron quantum dots. *Rep. Prog. Phys.* **64**, 701–736 (2001).
15. Weis, J., Haug, R. J., von Klitzing, K. & Ploog, K. Lateral transport through a single quantum dot with a magnetic field parallel to the current. *Surf. Sci.* **305**, 664–668 (1994).
16. Kouwenhoven, L. P. *et al.* Excitation spectra of circular, few-electron quantum dots. *Science* **278**, 1788–1792 (1997).
17. Ciorga, M. *et al.* Readout of a single electron spin based quantum bit by current detection. *Physica E* **11**, 35–40 (2001).
18. Fujisawa, T., Austing, D. G., Tokura, Y., Hirayama, Y. & Tarucha, S. Allowed and forbidden transitions in artificial hydrogen and helium atoms. *Nature* **419**, 278–281 (2002).
19. Hanson, R. *et al.* Zeeman energy and spin relaxation in a one-electron quantum dot. *Phys. Rev. Lett.* **91**, 196802 (2003).
20. Folk, J. A., Potok, R. M., Marcus, C. M. & Umansky, V. A gate-controlled bidirectional spin filter using quantum coherence. *Science* **299**, 679–682 (2003).
21. Lu, W., Ji, Z., Pfeiffer, L., West, K. W. & Rimbberg, A. J. Real-time detection of electron tunnelling in a quantum dot. *Nature* **423**, 422–425 (2003).
22. Fujisawa, T., Hayashi, T., Hirayama, Y., Cheong, H. D. & Jeong, Y. H. Electron counting of single-electron tunnelling current. *Appl. Phys. Lett.* **84**, 2343–2345 (2004).
23. Elzerman, J. M. *et al.* Few-electron quantum dot circuit with integrated charge read out. *Phys. Rev. B* **67**, 161308 (2003).
24. Elzerman, J. M., Hanson, R., Willems van Beveren, L. H., Vandersypen, L. M. K. & Kouwenhoven, L. P. Excited-state spectroscopy on a nearly-closed quantum dot via charge detection. *Appl. Phys. Lett.* **84**, 4617–4619 (2004).
25. Field, M. *et al.* Measurements of Coulomb blockade with a noninvasive voltage probe. *Phys. Rev. Lett.* **70**, 1311–1314 (1993).
26. Hanson, R. *et al.* Semiconductor few-electron quantum dot operated as a bipolar spin filter. Preprint at (<http://xxx.lanl.gov/abs/cond-mat/0311414>) (2003).
27. Khaetskii, A. V. & Nazarov, Y. V. Spin-flip transitions between Zeeman sublevels in semiconductor quantum dots. *Phys. Rev. B* **64**, 125316 (2001).

28. Golovach, V. N., Khaetskii, A. & Loss, D. Phonon-induced decay of the electron spin in quantum dots. Preprint at (<http://xxx.lanl.gov/abs/cond-mat/0310655>) (2003).
29. Woods, L. M., Reinecke, T. L. & Lyanda-Geller, Y. Spin relaxation in quantum dots. *Phys. Rev. B* **66**, 161318 (2002).
30. Erlingsson, S. I. & Nazarov, Y. V. Hyperfine-mediated transitions between a Zeeman split doublet in GaAs quantum dots: The role of the internal field. *Phys. Rev. B* **66**, 155327 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. P. DiVincenzo, H. A. Engel, T. Fujisawa, V. Golovach, Y. Hirayama, D. Loss, T. Saku, R. Schouten and S. Tarucha for technical support and discussions. This work was supported by the DARPA-QUIST programme, the ONR, the EU-RTN network on spintronics and the Dutch Organisation for Fundamental Research on Matter (FOM).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.M.K.V. (lieven@qt.tn.tudelft.nl).

Electrical detection of the spin resonance of a single electron in a silicon field-effect transistor

M. Xiao¹, I. Martin², E. Yablonovitch³ & H. W. Jiang¹

¹Department of Physics and Astronomy, University of California, Los Angeles, California 90095, USA

²Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

³Department of Electrical Engineering, University of California, Los Angeles, California 90095-1594, USA

The ability to manipulate and monitor a single-electron spin using electron spin resonance is a long-sought goal. Such control would be invaluable for nanoscopic spin electronics, quantum information processing using individual electron spin qubits and magnetic resonance imaging of single molecules. There have been several examples^{1,2} of magnetic resonance detection of a single-electron spin in solids. Spin resonance of a nitrogen-vacancy defect centre in diamond has been detected optically³, and spin precession of a localized electron spin on a surface was detected^{4,5} using scanning tunnelling microscopy. Spins in semiconductors are particularly attractive for study because of their very long decoherence times⁶. Here we demonstrate electrical sensing of the magnetic resonance spin-flips of a single electron paramagnetic spin centre, formed by a defect in the gate oxide of a standard silicon transistor. The spin orientation is converted to electric charge, which we measure as a change in the source/drain channel current. Our set-up may facilitate the direct study of the physics of spin decoherence, and has the practical advantage of being composed of test transistors in a conventional, commercial, silicon integrated circuit. It is well known from the rich literature of magnetic resonance studies that there sometimes exist structural paramagnetic defects⁷ near the Si/SiO₂ interface. For a small transistor, there might be only one isolated trap state that is within a tunnelling distance of the channel, and that has a charging energy close to the Fermi level.

When a defect is present, the source/drain channel current can experience random telegraph signal (RTS), jumping between two discrete current values. These arise from two possible trapped electric charge states of the defect. The two charge states can correspond to the two spin orientations of a trapped electron. Field effect transistor (FET) current senses electrostatic charge (by definition), and can thus sense single-electron spin resonance.

In view of the high degree of perfection of silicon devices, 80% of the transistors that we tested actually had no such trap states at all. In those cases we applied a high voltage spike to the gate in an attempt to induce a paramagnetic defect for study. Electron spin in low-transition-temperature semiconductors is now recognized⁸ to have considerable potential for storing and manipulating quantum information. Some proposed qubit architectures trap electron spin in a controlled potential well⁹ at the Si-Ge interface¹⁰ and use transistor-like structures¹¹ for sensing.

Figure 1a sketches a simplified version of the transistor device in our experiment. The essential requirement for detecting a single spin flip is the conversion of spin-orientation to electric charge. In our approach, the Fermi level is adjusted so that it lies between the upper and lower Zeeman levels as illustrated in Fig. 1b. If the lower Zeeman level is occupied by one electron, as in Fig. 1b, it cannot accept any additional electrons from the Fermi level. If only the upper Zeeman level is occupied, as in Fig. 1c, then an additional electron can be transferred from the Fermi sea to the lower Zeeman level. The distinction between two trapped charges $2e^-$ (in Fig. 1c) versus one trapped charge $1e^-$ (in Fig. 1b) can be sensed, because an FET is basically an electrometer. Thus spin orientation is converted to electric charge, which is sensed by the source/drain current in an FET transistor.

Our n-channel Si transistors are operated near pinch-off, where the channel size is roughly 300 nm wide $\times$ 240 nm long. The signature of a single trap state is shown in the RTS source/drain current of Fig. 2a. Superimposed on the monotonically increasing background source/drain current is stochastic switching between two discrete values of channel current. This switching is the well-known RTS (for a comprehensive review, see ref. 12) that is a hallmark of the capture and emission of one electron by a single trap state. The well-defined RTS evolution in Fig. 2a demonstrates that over the 690 to 740 mV range, the trap is energetically well isolated from other traps.

A filled trap implies electrostatic repulsion that diminishes the channel current. At high gate voltages (near point A in Fig. 2a), the Fermi level E_F is well above the trap level E_T . Thus the trap is almost always filled, repelling electrons and allowing less current to flow in the source/drain channel. In contrast, at low gate voltages (near point C in Fig. 2a) when E_F is well below E_T , the trap is empty most of the time and the high current state is more probable. At the midpoint, when $E_F \approx E_T$ (near point B in Fig. 2a) the probability of the trap filling is about 50%. Thus the source/drain current senses the charge state of the trap.

The trap studied here is a very stable defect because the behaviour is reproducible over many thermal cycles from room temperature to

liquid He³ temperatures. From the signal-to-noise ratio, we find that the charge sensitivity of the our data acquisition system is $\sim 10^{-4} e \text{ Hz}^{-1/2}$.

The Zeeman shift of the single trap¹³ can readily be identified by studying the trap energy shift of the point at which the probability of the trap filling is 50% (where E_F lines up with the trap energy E_T), as a function of magnetic field. Figure 2b shows the Zeeman shift of this point as a function of an in-plane magnetic field. The trap energy shift was inferred from the gate voltage shift by using the procedure in ref. 13.

On the basis of the sign of the Zeeman shift, ref. 13 shows that the charging transition transfers from a single-charge state $1e^-$ to a double-charge state $2e^-$; that is, the charging is $1 \rightarrow 2$ rather than $0 \rightarrow 1$. In the energy diagram of Fig. 1b, the empty trap is modelled as an unpaired electron (for example, a dangling bond) that occupies the level E_T (the central red dashed line). In the presence of the magnetic field B , the single-electron state undergoes Zeeman splitting, indicated by the two solid lines at energies $E_T \pm (1/2)E_Z$. At low temperatures and high fields, only the lower spin state is occupied. If E_F is raised, an additional electron from the channel can tunnel into the upper spin state in Fig. 1b, forming a two-electron singlet state (for example, a lone pair). Thus the Fermi energy required for forming the two-electron state would increase when B is increased, as suggested by Fig. 2b. In contrast, an initially 'spinless' empty trap would fill the lower Zeeman level, producing the opposite field dependence—that is, the required Fermi energy would decrease with increasing B , contrary to observation. Therefore the initial empty trap begins in a $1e^-$ paramagnetic state ($S = 1/2$) (high current state) while the filled trap (lower current state) is a $2e^-$ singlet state.

Our electron spin resonance (ESR) detection scheme is based on the changing balance between the two source/drain current states of the transistor, when the Larmor precession frequency produces spin flips. In effect, this is transistor-current-detected ESR, as illustrated in Fig. 1c. When the paramagnetic spin flips, the lower Zeeman level becomes available for trapping an additional electron. The trapping event diminishes the average source/drain current. A rate equation analysis of this trap/channel configuration gave¹⁴ the ESR-induced change in trap-filling probability.

To detect the ESR-microwave-induced change, we measured channel current at a fixed microwave frequency for 300 ms, during which there are about a few thousand RTS switching events, yielding good statistics for the average source/drain current. Figure 3a represents a fragment of such a trace over a 1 ms time interval. To complete the current versus magnetic field dependence, full 300-ms traces are taken at 150–250 different magnetic fields.

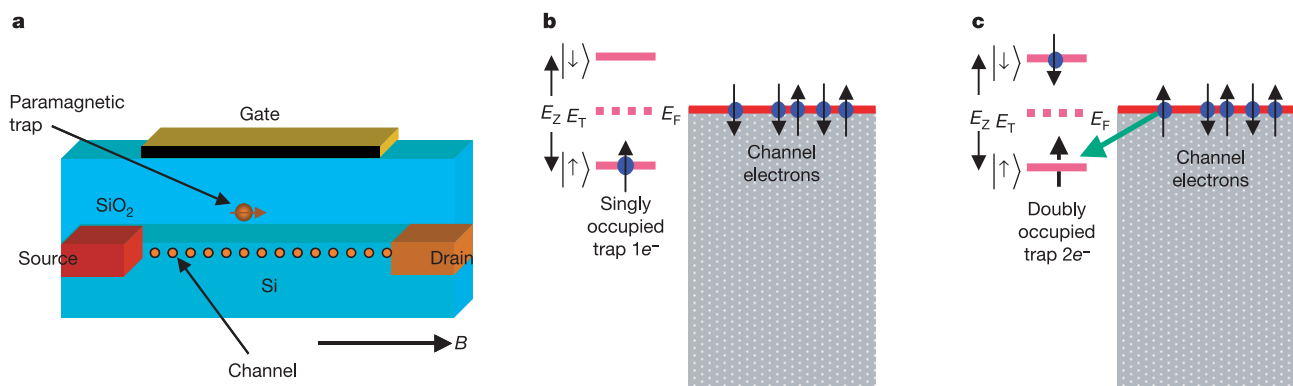


Figure 1 The mechanism for spin-to-charge conversion. **a**, A single paramagnetic trap, situated in a standard commercial n-channel Si FET that constitutes our sample. **b**, The Zeeman-split trap level relative to the FET channel Fermi level. The Fermi level would have to shift towards the upper Zeeman level to reach 50% occupation probability. (The singly

occupied state should actually be down-shifted by the Coulomb correlation energy U , not shown.) **c**, If the spin flips under ESR, the lower Zeeman level can become filled, producing the doubly occupied trap.

Because the signals are sometimes noisy, a systematic statistical procedure was used to measure the trap charge state. A histogram of the source/drain current data versus time was plotted as shown in Fig. 3c. In the histogram, the empty and filled trap states are represented by the large and small peak, respectively. For the perfect case of two discrete states, we expect two narrow peaks in the histogram. The broadening of the peaks in Fig. 3c is caused by noise. The charge trapping probability ratio is proportional to the area ratio of the two peaks. For certain traps, whose charge produces only a small change in source/drain channel current, an additional step was taken to avoid noise errors. A sophisticated algorithm¹⁵, for detection of abrupt step changes, was executed numerically. Figure 3a is the raw RTS, containing noise, whereas Fig. 3b shows the noiseless two-state switching, reconstructed by the algorithm.

We now present the ESR detection results for the single paramagnetic trap. Figure 4a shows the change in average source/drain current induced by a microwave frequency of 45.1 GHz. In Fig. 4a, an ESR peak in average current is centred around 16,025 G. Averaging blocks of four adjacent magnetic fields, the signal/noise ratio is >4:1, and the ESR feature is reproducible in different runs, and for different traps, in different samples.

We find that the ESR signal is most pronounced in the range of gate voltages corresponding to a paramagnetic (nearly empty) trap

(that is, between points B and C in Fig. 2a). This is consistent with our assignment of filled and empty trap states.

The ESR signature is only found at temperatures below about 1 K. At those temperatures, the electron magnetic moment is substantially polarized, and in any case, microwave heating limits the temperature. From the RTS Boltzmann occupation probability as a function of voltage¹³, we find that the effective temperature rises to about 1.5 K when a moderate microwave power of 0.1 mW is applied to the sample, even though the bath temperature still remains at about 0.4 K. The sensitivity of our time-averaged readout is thus limited by the microwave heating. From our signal-to-noise ratio, we estimate that our ESR signal could be read in a single shot, if the ESR were driven by a pulsed 180-GHz source. In the future, heating can be avoided by using *g*-tensor modulation^{16,17} for single spin rotation.

Similar runs have been carried out at other frequencies and in various samples. The ESR magnetic resonance as a function of microwave frequency is plotted in Fig. 4b. For $hf = g\mu_B B$, a *g*-factor of 2.02 ± 0.015 is obtained. For a *g*-factor reference we used both a bulk Si crystal doped with $2 \times 10^{18} \text{ cm}^{-3}$ phosphorous; and the channel electrons in our n-channel Si transistor (Fig. 1a). Both agreed with the well-established values of $g \approx 1.998$. Because conduction electrons always¹⁸ have $g < 2$, and paramagnetic centres in SiO₂ always⁷ have $g > 2$, our results indicate a paramagnetic centre in the oxide, or at the SiO₂/Si interface. Our observed *g*-value is somewhat larger than that for some known paramagnetic centres near the Si/SiO₂ interface. A P_b paramagnetic centre is known to

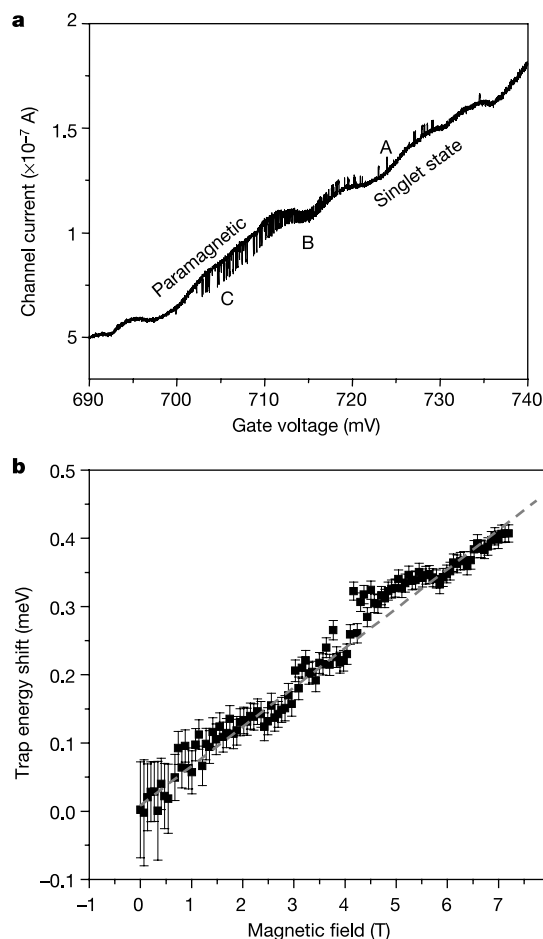


Figure 2 Random telegraph signal senses the defect energy level. **a**, The channel current I was recorded over a narrow gate voltage ramp from 690–740 mV, swept in a 30-ms time interval, with a 1-mV source/drain bias voltage at 1.3 K. The discrete upper and lower current states sense the two charge states, which can monitor the two electron spin states. **b**, The positive Zeeman shift of the trap energy versus magnetic field implies a $1e^-$ to $2e^-$ transition in the defect, rather than a $0e^-$ to $1e^-$ transition. The error bars indicate the standard deviation in the raw data.

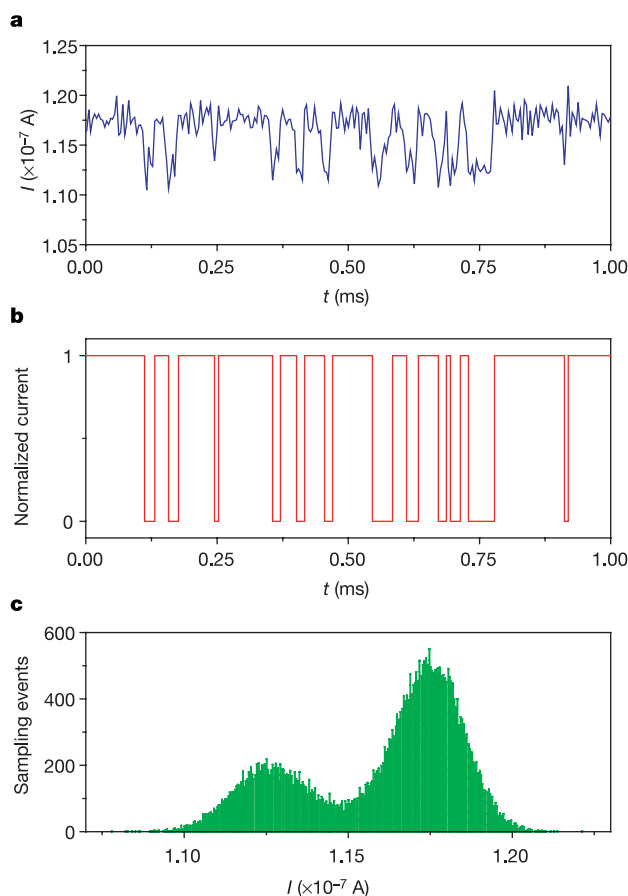


Figure 3 Detailed analysis of the random telegraph signal. **a**, The raw random telegraph data displayed for a time interval of 1 ms. **b**, An algorithm for detection of abrupt changes is applied to the raw data to reconstruct the two RTS levels. This procedure reduces the statistical errors due to noise. **c**, A histogram of the time-domain data. The large and small peaks represent the empty and filled trap states, respectively.

have a g -factor of 2.006 along the $\langle 100 \rangle$ direction, whereas the E' centre is expected to have $g = 2.001$. One possibility is that we are looking at a centre that has a local structure different from that of these two typical examples. Another possibility is that the low-

density conduction channel electrons might have slight ferromagnetic ordering, giving rise to a local field that slightly increases the apparent g -factor of the trap.

The full-width, half maximum linewidth of the observed resonance is about 5–10 G, which corresponds to the spin-spin relaxation time T_2 of about 0.1 μ s. This time is much less than the typical 100- μ s lifetime for isolated paramagnetic spins (the paramagnetic E' centre in irradiated fused quartz has a T_2 of about 25 μ s at room temperature; see for example, ref. 19) in SiO_2 . There might be additional decoherence caused by the conduction electrons in the channel, or it might be caused by power broadening. Indeed, our microwave power produces a Rabi frequency $\gamma H_1 \approx 10$ kHz, comparable to the trap tunnelling time of about 20 kHz, where γ is the gyromagnetic ratio and H_1 is the microwave B field. For actual quantum information processing, the source/drain channel should only be un-pinned at the end of a quantum computation for readout, well within the T_1 lifetime.

At the strong Rabi frequency of $\gamma H_1 \approx 10$ kHz that is present at the end plate of our rectangular waveguide, nonlinear effects emerge. At lower microwave powers $\gamma H_1 \approx 1$ kHz, the single-occupancy trap probability is correctly observed to diminish under ESR, as expected from the reasoning in Fig. 1c. However at higher powers, $\gamma H_1 \approx 10$ kHz, the ESR-induced-current signal inverts, leading to an increase in occupancy trap probability, as plotted in Fig. 4a. We plotted the signal at high microwave power because it has a better signal-to-noise ratio.

We also see a nonlinear ESR response in the tunnelling dynamics. As shown in Fig. 5, the frequency of RTS jumps, which is proportional to the tunnelling rate, can change as much as 10% at the ESR condition. At low microwave power, $\gamma H_1 \approx 1$ kHz, the tunnelling rate increases at resonance, but at high Rabi frequency, $\gamma H_1 \approx 10$ kHz, the tunnelling rate inverts, as shown in Fig. 5. These inverted signals in both the source/drain current, and the tunnelling frequency, are only observed for high Rabi frequencies that are comparable to the tunnelling frequency. The nonlinear ESR response may be partly associated with non-resonant microwave-induced changes in the tunnelling rate, but we have not yet identified the dominant nonlinear mechanism.

Our evidence in support of the ESR signal arising from a single electron is summarized as follows: (1) The ESR signature is only observed in the random telegraph signal crossover region, predominantly between points B and C of Fig. 2a. (2) The random telegraph signal switching statistics, and thermal occupation probabilities, correspond to those of a single defect. (3) The g -factor value rules out channel electrons. The experiment demonstrates that an FET channel can effectively monitor the magnetic resonance of an adjacent single spin. $\square$

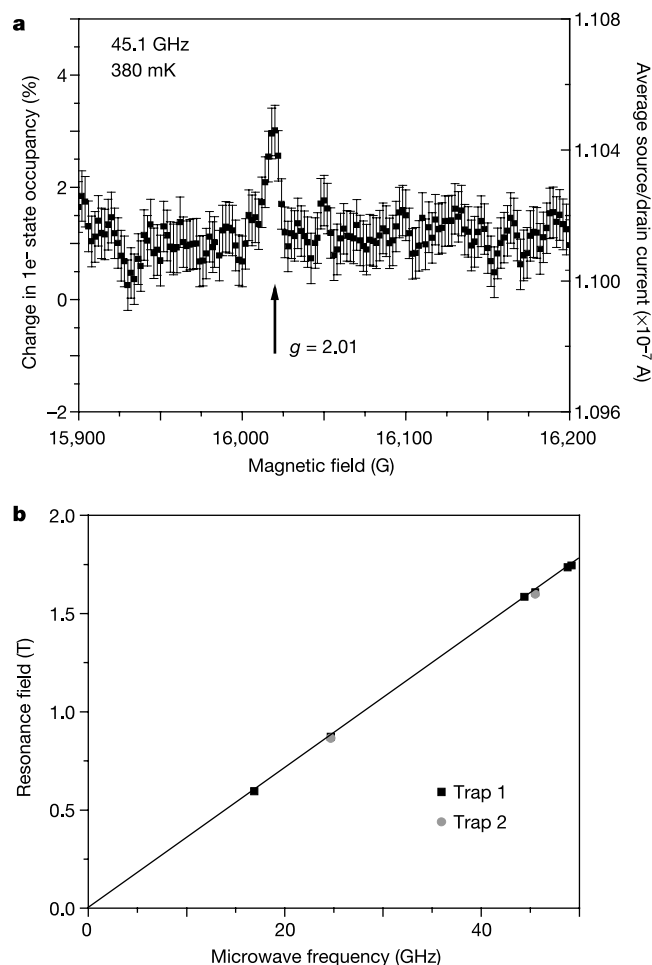


Figure 4 The single electron ESR signal. **a**, The single-occupancy trap probability, and the average source/drain current, versus magnetic field for a fixed microwave frequency. The peak represents electron spin resonance. The error bars (about 1%) indicate the standard deviation in a 300-ms data set averaged over four adjacent magnetic fields. **b**, The linear relationship of resonant magnetic field versus microwave frequency indicates a g -factor of 2.02 ± 0.015 .

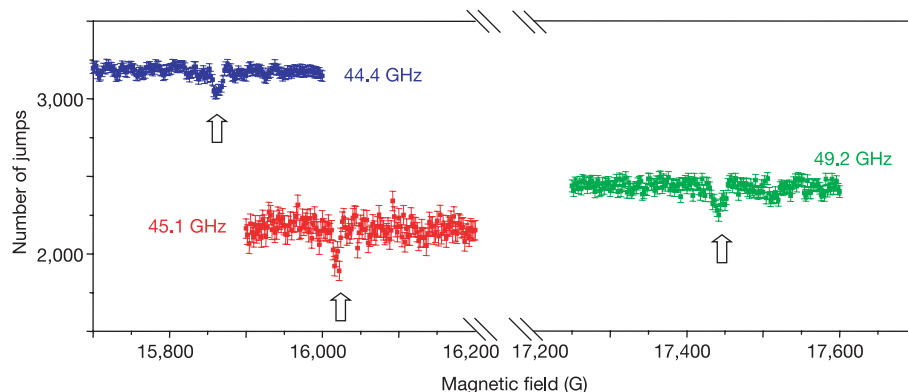


Figure 5 ESR induced change in tunnelling rates. The tunnelling rate between charge states is also substantially modified ($\sim 10\%$ change) at ESR. The number of transitions, over a 300-ms time interval, between the paramagnetic and singlet states is plotted as a

function of field for three different microwave frequencies. The resonance positions are marked by the vertical arrows. The error bars represent the standard deviation in the given data set.

Received 5 April; accepted 7 June 2004; doi:10.1038/nature02727.

- Köhler, J. *et al.* Magnetic resonance of a single molecular spin. *Nature* **363**, 242–244 (1993).
- Wrachtrup, J., von Borczyskowski, C., Bernard, J., Orrit, M. & Brown, R. Optical detection of magnetic resonance in a single molecule. *Nature* **363**, 244–245 (1993).
- Jelesko, F., Gaebel, T., Popa, I., Gruber, A. & Wrachtrup, J. Observation of coherent oscillations in a single electron spin. *Phys. Rev. Lett.* **92**, 076401 (2004).
- Manassen, Y., Hames, R. J., Demuth, J. E. & Castellano, A. J. Direct observation of the precession of individual paramagnetic spins on oxidized silicon surfaces. *Phys. Rev. Lett.* **62**, 2531–2534 (1989).
- Durkan, C. & Welland, M. E. Electronic spin detection in molecules using scanning-tunneling-microscopy-assisted electron-spin resonance. *Appl. Phys. Lett.* **80**, 458–460 (2002).
- Tyryshkin, A. M., Lyon, S. A., Astashkin, A. V. & Raitsimring, A. M. Electron spin relaxation times of phosphorus donors in silicon. *Phys. Rev. B* **68**, 193207 (2003).
- Poindexter, E. H. in *Semiconductor Interfaces, Microstructure, and Devices* (ed. Feng, Z. C.) Ch. 10, 229–256 (Inst. Phys. Publ., Bristol, Philadelphia, 1993).
- Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
- Friesen, M. *et al.* Practical design and simulation of silicon-based quantum-dot qubits. *Phys. Rev. B* **67**, 121301 (2003).
- Vrijen, R. *et al.* Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures. *Phys. Rev. A* **62**, 012306 (2000).
- Kane, B. E. A silicon based nuclear spin quantum computer. *Nature* **393**, 133–137 (1998).
- Kirton, M. J. & Uren, M. J. Noise in solid-state microstructures: a new perspective on individual defects, interface states and low-frequency (1/f) noise. *Adv. Phys.* **38**, 367–468 (1989).
- Xiao, M., Martin, I. & Jiang, H. W. Probing the spin state of a single electron trap by random telegraph signal. *Phys. Rev. Lett.* **91**, 078301 (2003).
- Martin, I., Mozyrsky, D. & Jiang, H. W. A scheme for electrical detection of single electron spin resonance. *Phys. Rev. Lett.* **90**, 018301 (2003).
- Liu, W., Ji, Z., Pfeiffer, L., West, K. W. & Rimbarg, A. J. Real-time detection of electron tunnelling in a quantum dot. *Nature* **423**, 422–425 (2003).
- Kato, Y. *et al.* Gigahertz electron spin manipulation using voltage-controlled g-tensor modulation. *Science* **299**, 1201–1204 (2003).
- Yablonovitch, E. *et al.* Optoelectronic quantum telecommunications based on spins in semiconductors. *Proc. IEEE* **91**, 761–780 (2003).
- Wallace, W. J. & Silsbee, R. Spin resonance of inversion-layer electrons in silicon. *Phys. Rev. B* **44**, 12964–12968 (1991).
- Eton, S. S. & Eaton, G. R. Irradiated fused-quartz standard sample for time-domain EPR. *J. Magn. Res. Ser. A* **102**, 354–356 (1993).

Acknowledgements We would like to thank P. M. Lenahan for information regarding the relaxation of spin centres in SiO₂, and K. Wang for helping to secure the samples. The work was supported by the Defense Advanced Research Projects Agency and the Defense MicroElectronics Activity.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.W.J. (jiangh@physics.ucla.edu).

Surface transfer doping of diamond

P. Strobel, M. Riedel, J. Ristein & L. Ley

Institute for Technical Physics, University of Erlangen, 91054 Erlangen, Germany

The electronic properties of many materials can be controlled by introducing appropriate impurities into the bulk crystal lattice in a process known as doping. In this way, diamond (a well-known insulator) can be transformed into a semiconductor¹, and recent progress in thin-film diamond synthesis has sparked interest in the potential applications of semiconducting diamond^{2,3}. However, the high dopant activation energies (in excess of 0.36 eV) and the limitation of donor incorporation to (111) growth facets only have hampered the development of diamond-based devices. Here we report a doping mechanism for diamond, using a method that does not require the introduction of foreign atoms into the diamond lattice. Instead, C₆₀ molecules are evaporated onto the hydrogen-terminated diamond surface, where they induce a sub-surface hole accumulation and a significant rise in two-dimensional conductivity. Our observations bear a resemblance to the so-called surface conductivity of diamond^{4–8} seen when hydrogenated diamond surfaces are exposed to air, and support an electro-

chemical model in which the reduction of hydrated protons in an aqueous surface layer gives rise to a hole accumulation layer^{6,7}. We expect that transfer doping by C₆₀ will open a broad vista of possible semiconductor applications for diamond.

The C₆₀ was evaporated onto differently prepared diamond surfaces and onto a silica glass substrate. After each evaporation step the conductance of the respective sample was measured as described below and without breaking the vacuum. The closed symbols in Fig. 1 refer to the conductance of two different homo-epitaxial diamond samples with hydrogenated surfaces as a function of C₆₀ coverage in a double-logarithmic plot. The starting point is in both cases the annealed state (see Methods) with a conductance well below 10^{−12} S (leftmost data point). With increasing C₆₀ coverage the conductance rises rapidly by more than six orders of magnitude and saturates at 10^{−6} S for sample D20 and 10^{−5} S for sample D34 at a C₆₀ coverage between 4 and 8 monolayers (ML). Two control experiments (open symbols) were performed with silica glass and with diamond sample D20, this time with an oxidized instead of a hydrogenated surface as the substrate for C₆₀ evaporation. The negative results presented in Fig. 1 confirm that the conductivity necessarily requires the combination of C₆₀ and hydrogenated diamond; the C₆₀ layer alone on an insulating substrate (glass) or in combination with an oxidized diamond surface is not sufficient.

We propose that the cause of the observed conductivity is a transfer doping mechanism whereby electrons are transferred from diamond to the C₆₀ layer such that an equal amount of electrons and holes reside on the C₆₀ and diamond side of the interface, respectively. For this to work, the energy level scheme as sketched in Fig. 2a has to be considered. The left-hand part depicts the valence and conduction band edges (VBM and CBM) of diamond, together with the Fermi level *E_F*. On the right-hand side a limited number of occupied and unoccupied molecular orbitals of C₆₀ are schematically sketched with the highest occupied (HOMO) and the lowest unoccupied (LUMO) bracketing the fundamental gap of C₆₀. The common reference level is the vacuum level *E_{VAC}*. For hydrogen-terminated diamond the electron affinity is negative, that is, the CBM lies 1.3 eV above *E_{VAC}* (refs 9, 10). The electron affinity *E_{VAC}* − *E_{LUMO}* of C₆₀ is measured to be about 2.7 eV for the C₆₀ molecule¹¹, and this places the Fermi level of C₆₀, which lies between the HOMO and the LUMO, well below the *E_F* of diamond. Hence,

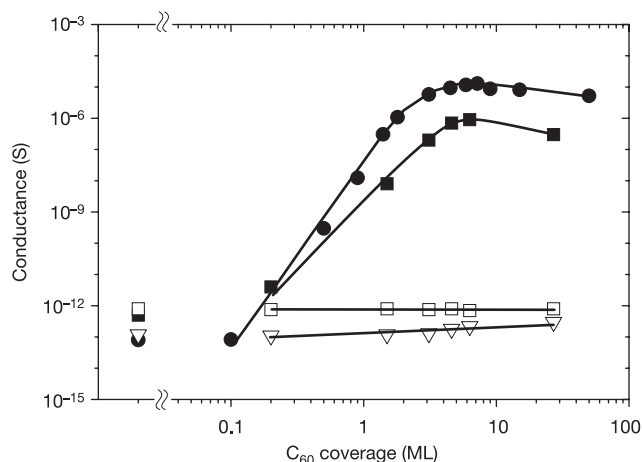


Figure 1 Conductance of different substrates upon evaporation of C₆₀ in ultrahigh vacuum. Filled circles and squares, hydrogen-terminated diamond, samples D20 and D34, respectively. Open squares, oxygen-terminated diamond; open triangles, silica glass. The leftmost data points correspond to the clean surfaces after mild annealing at about 250 °C. Note the logarithmic scales of the plot. The C₆₀ coverage is given in ML; 1 ML corresponds to an areal density of 1.15 × 10¹⁴ C₆₀ molecules per cm². The lines through the data points are guides to the eye.

an electron flow from the diamond valence band enters the LUMO of C_{60} upon contact. The holes left behind set up a space charge region in diamond that results in an upward band bending, as indicated schematically in Fig. 2b. At the same time the Fermi level of the C_{60} layer moves towards E_{LUMO} and the charge exchange comes to a halt when the two Fermi levels line up. The amount of band bending and the Fermi level shift in C_{60} are connected through the condition of charge neutrality: the number of holes in diamond equals that of electrons in C_{60} . Combining Poisson's equation for the space charge density on the diamond side and the Fermi occupation statistics for the diamond and the C_{60} states, it is a somewhat lengthy but straightforward task to show that the equilibrium charge concentration is entirely determined by the areal density of C_{60} molecules and the energy separation $\Delta = E_{LUMO,C60} - VBM_{diamond}$, which is a fixed quantity independent of any charge exchange and band bending. The carrier density can be estimated from the measured conductivity provided we know the mobility of the respective carriers. For solid C_{60} the electron mobility at 300 K does not exceed $0.1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ (ref. 12), whereas the mobility of surface-near holes in diamond lies typically between 50 and $100 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ (ref. 13). Hence the conductivity must be ascribed to the holes in diamond. Adopting an average value of $\mu \approx 70 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ for the hole mobility we translate the two-dimensional conductivity $\sigma_{\square} = e\mu n$ (where e is the elementary charge) into the two-dimensional hole density n (right-hand scale in Fig. 3). For sample D34 we estimate a maximum sheet carrier concentration of $1.5 \times 10^{12} \text{ cm}^{-2}$ for a C_{60} coverage of 7 ML. The simulation based on Poisson's equation and Fermi statistics yields a corresponding value of 0.24 eV for Δ . With this value of Δ the concentration of transferred electrons and the ensuing conductivity have been calculated as a function of C_{60} coverage. The result, compared in Fig. 3 with the measurements for sample D34, is far off the mark. With decreasing coverage, the actual conductance drops precipitously by six orders of magnitude between about 2 and 0.1 ML of C_{60} , whereas the calculated value decreases by a mere factor of ten over the same range. The straightforward solution, namely to attribute the experimental drop to a corresponding drop in mobility, is unreasonable. If anything, mobilities tend to increase with decreasing carrier density. This leaves an increase in Δ as the only cause for the discrepancy in the framework of the model. In Fig. 4 we have calculated values for Δ as a function of C_{60} coverage such that they reproduce the conductivity data of sample D34. A

very similar relationship, albeit with fewer points, holds for D20. In essence, Fig. 4 says that it costs 0.24 eV to transfer an electron from the diamond valence band maximum into the LUMO of a C_{60} layer several ML thick, whereas the same transfer costs 1.55 eV if it takes place into an isolated C_{60} molecule. Phrased differently, the electron affinity increases by 1.31 eV when going from a single C_{60} molecule to a C_{60} solid.

The electron affinity is, strictly speaking, the difference of two total energies, namely that of the system in its N -electron ground state $E(N)$ and that of the $N + 1$ electron system also in its ground state, $E(N + 1)$. The difference $E(N) - E(N + 1)$ is a true many-electron quantity that depends on the size of the system. For the sake of argument it is convenient to split the many-electron expression into a difference of one electron energies (that is, $E_{VAC} - E_{LUMO}$) and a many-body correction that is usually termed the effective correlation energy U . This energy involves the Coulomb interaction of the additional electron brought into the lowest accessible electronic state of the system with all the other electrons already present. As such, it depends on, and is reduced by, the relaxation of the electrons already present, and this relaxation in turn depends strongly on the polarizability of the system. Consequently, the effective correlation energy is significantly lower for an electron brought to the conduction band minimum of C_{60} in the solid phase than into the LUMO of an isolated C_{60} molecule. For such a molecule, theory gives values between 2.7 and 3.1 eV for the correlation energy, whereas an experimental estimate yields a slightly larger value of 3.3 eV (ref. 14 and references therein). The same quantity for the solid state was found between 0.8 and 1.3 eV theoretically and between 1.4 and 1.6 eV experimentally¹⁴. The difference $\delta U \approx 1.8 \pm 0.2 \text{ eV}$ in effective correlation energy is in reasonable agreement with the change of 1.31 eV in Δ derived above. In fact, electrons and holes in the transfer doping process have only to be separated across the width of the diamond space charge layer plus the C_{60} layer thickness. This leaves a finite—though hard to specify—Coulomb energy by which the 'band offset' energy Δ should be smaller than the difference between the diamond ionization potential and the C_{60} electron affinity. The remaining Coulomb energy is expected to decrease with increasing C_{60} layer thickness, and the decrease in Δ should indeed be smaller than δU by this difference.

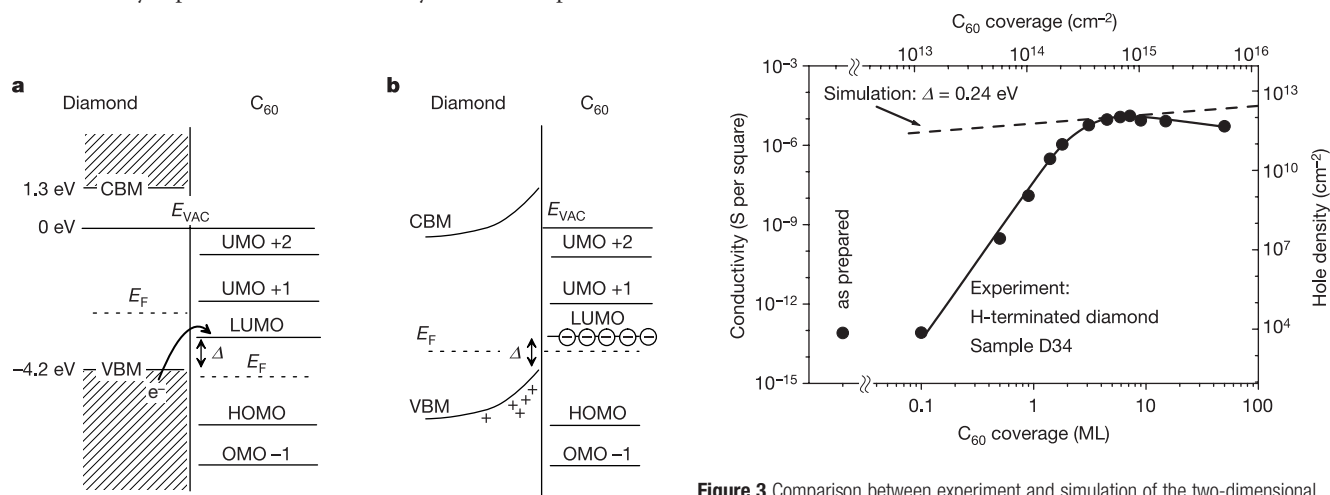


Figure 2 Schematic representation of the surface transfer doping of intrinsic diamond by C_{60} . **a**, Level scheme before electron transfer from the diamond side with the higher-lying E_F to the C_{60} side with the lower-lying E_F . **b**, Band bending in chemical equilibrium. An equal amount of holes in diamond and electrons in the lowest (normally, that is for neutral C_{60} at zero temperature) unoccupied molecular orbital (LUMO) of the C_{60} layer with a concomitant electrostatic potential on the diamond side is established.

Figure 3 Comparison between experiment and simulation of the two-dimensional conductivity of hydrogen-terminated diamond as a function of C_{60} coverage. The simulation is based on a quantitative evaluation of the surface transfer doping model illustrated in Fig. 2. A 'band offset' of $\Delta = 0.24 \text{ eV}$, independent of C_{60} coverage, has been assumed as the only free parameter of the simulation. The sheet conductivity (left-hand ordinate) and the areal hole density (right-hand ordinate) are proportional to each other with the hole mobility of $70 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and the elementary charge as proportionality constants.

The fact that no conductivity is observed for an oxidized diamond surface stresses the importance of the low ionization energy of diamond in conjunction with the rather high electron affinity of C_{60} . Removal of hydrogen raises the ionization potential by 1.6 eV and oxidation by more than 3 eV (ref. 10), with the result that transfer doping is energetically not possible and conductivity no longer observed. As a final control experiment we used hydrogenated but nitrogen-doped synthetic type Ib diamond as substrate and observed no conductivity ($\sigma_{\square} \leq 10^{13}$ S per square) upon evaporation of C_{60} . The electron affinity as a pure surface property does not depend on substrate doping. Thus the same transfer of electrons to C_{60} as for intrinsic diamond is expected. However, nitrogen forms a deep donor in diamond and all holes that are created as a result of transfer doping recombine with electrons from these compensating defects. The charge exchange results in a positive space charge (and a corresponding upward band bending) in the diamond; this time, however, the positive space charge is localized as ionized N^+ defects and not present as mobile holes. On the C_{60} side we expect no difference compared to the experiment on intrinsic diamond; in particular, an electron transfer into the LUMO is also expected to take place as well. The low conductance of the sample confirms the negligible mobility of those electrons and is an indirect proof that the assignment of the observed conductivity to holes in diamond is correct.

The results presented above bear a very close resemblance to the so-called surface conductivity of diamond. Discussions on the origin of this type of surface conductivity, which is observed when hydrogenated diamond surfaces are exposed to air, have generated controversy. The suggestions range from hydrogen-induced acceptor states⁴, eventually in combination with compensating surface defects⁵, to an electrochemical model in which the reduction of hydrated protons in an aqueous surface layer gives rise to the hole accumulation layer^{6,7}. The electrochemical model met with a certain amount of scepticism because a charge transfer from diamond to an adsorbate across the interface was thought unrealistic. In view of the results presented here it can no longer be doubted that the electrochemical model for the air induced surface conductivity is basically correct, because it relies on a mechanism similar to the one described here. From a technical point of view the new transfer doping by C_{60} is likely to replace the air-induced surface conductivity for most of the devices that have been realized on the basis of the surface conductivity of diamond^{8,15}. The ease of fabrication, higher reliability and reproducibility of the C_{60} doping are assets that will ultimately improve the device characteristics and could bring new concepts to fruition. □

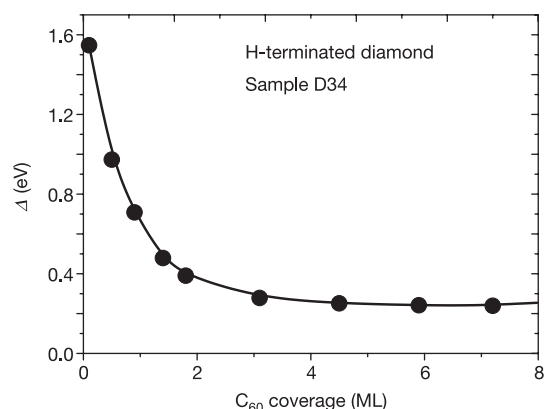


Figure 4 Reduction of the 'band offset' Δ (see Fig. 2) between the diamond valence band maximum and the C_{60} LUMO with increasing C_{60} coverage. This reduction reflects the increase of the electron affinity of C_{60} when going from isolated molecules to the molecular solid.

Methods

Sample preparation and characterization

For the experiments the following types of diamond samples have been used. The main work was performed on undoped, homoepitaxial diamond layers with (100) orientation (samples D20 and D34) that were grown by microwave-assisted vapour phase epitaxy to a thickness of about 0.5 μm on nitrogen-doped (N concentration 10^{19} cm^{-3}), non-conducting, synthetic high-pressure, high-temperature diamond substrates. A natural, boron-doped and therefore conducting, Ila diamond was used for photoelectron spectroscopy of the C_{60} layer, and finally a nitrogen-doped substrate (without an intrinsic epilayer on top) was used for reference measurements. All samples, except those for which the converse is explicitly stated, were plasma-hydrogenated to saturation in a separate preparation chamber before the experiments.

The experiments were performed in interconnected ultrahigh vacuum chambers with a pressure below 10^{-9} mbar. Owing to their contact with air, the as-prepared samples exhibited the well-known atmosphere-induced surface conductivity, which had to be removed in a first preparation step. To this end, the diamond samples were initially annealed in UHV at temperatures between 200 and 300 °C. These temperatures are known to remove all residual surface conductivity while leaving the hydrogen termination intact¹⁶. This was confirmed *in situ* by total photoelectron yield spectra which exhibit the signature characteristic for hydrogen-terminated surfaces with negative electron affinity¹⁷.

The C_{60} was evaporated onto the diamond surfaces and onto the silica glass reference substrate at a rate of typically 0.1 ML min^{-1} . The nature of the evaporated layer as C_{60} was confirmed on the basis of its valence band spectrum, the characteristic electron energy loss spectrum, and the shift of 1.3 eV of its $C1s$ line relative to the $C1s$ binding energy of diamond—all measured by photoelectron spectroscopy¹⁸. Photoelectron spectroscopy was also used to determine the C_{60} layer thickness from the attenuation of the diamond signal.

Conductivity measurements

After each evaporation step of C_{60} , two gold-coated, spring-loaded tips of $r = 0.05$ mm radius at a centre distance of $D = 2$ mm were pressed onto the sample surface; the conductance of the different specimens was then determined *in situ* by measuring the $I - V$ characteristics between -10 and $+10$ V. The ohmic behaviour was verified for each measurement and the conductance derived by a linear regression. The conductance is (aside from a geometric factor γ of the order of unity) equal to the two-dimensional surface conductivity $\sigma_{\square}$. For the tip arrangement used here, $\gamma = \ln(D/r - 1)/\pi = 1.17$, and in view of the large variation of conductivities observed in our experiments, conductance and two-dimensional conductivity could be used as synonyms throughout this paper.

Received 8 April; accepted 8 June 2004; doi:10.1038/nature02751.

- Nebel, C. E. & Ristein, J. (eds) *Thin-Film Diamond I Monogr. Ser. Semiconductors and Semimetals* Vol. 76, 145–259 (Elsevier, New York, 2003).
- Nebel, C. E. & Ristein, J. (eds) *Thin-Film Diamond II Monogr. Ser. Semiconductors and Semimetals* Vol. 77, 97–338 (Elsevier, New York, 2004).
- Isberg, J. et al. High carrier mobility on single-crystal plasma deposited diamond. *Science* **297**, 1670–1672 (2002).
- Denisenko, A. et al. Hypothesis on the conductivity mechanism in hydrogen terminated diamond. *Diamond Relat. Mater.* **9**, 1138–1142 (2000).
- Takeuchi, D., Yamanaka, S. & Okushi, H. Schottky junction properties of the high conductivity layer of diamond. *Diamond Relat. Mater.* **11**, 355–358 (2002).
- Ri, S. G. et al. Formation mechanism of p-type surface conductive layer on deposited diamond films. *Jpn. J. Appl. Phys.* **34**, 5550–5555 (1995).
- Maier, F., Riedel, M., Mantel, B., Ristein, J. & Ley, L. The origin of surface conductivity in diamond. *Phys. Rev. Lett.* **85**, 3472–3475 (2000).
- Kawarada, H. Hydrogen terminated diamond surfaces and interfaces. *Surf. Sci. Rep.* **26**, 205–259 (1996).
- Cui, J. B., Ristein, J. & Ley, L. The electron affinity of the bare and hydrogen covered single crystal diamond (111) surface. *Phys. Rev. Lett.* **81**, 429–432 (1998).
- Maier, F., Ristein, J. & Ley, L. Electron affinity of plasma-hydrogenated and chemically oxidized diamond (100) surfaces. *Phys. Rev. B* **64**, 165411 (2001).
- Yang, S. H., Petiette, C. L., Conceicao, J., Cheshnovsky, O. & Smalley, R. F. UPS of buckminsterfullerene and other large clusters of carbon. *Chem. Phys. Lett.* **139**, 233–238 (1987).
- Haddon, R. C. et al. C_{60} thin film transistors. *Appl. Phys. Lett.* **67**, 121–123 (1995).
- Hayashi, K. et al. Investigation of the effect of hydrogen on electrical and optical properties in chemical vapor deposited homoepitaxial diamond films. *J. Appl. Phys.* **81**, 744–753 (1997).
- Gunnarsson, O. Superconductivity in fullerenes. *Rev. Mod. Phys.* **69**, 575–606 (1997).
- Gluche, P., Aleksov, A., Vescan, A., Ebert, W. & Kohn, E. Diamond surface-channel FET structure with 200 V breakdown voltage. *IEEE Electr. Dev. Lett.* **18**, 547549 (1997).
- Riedel, M., Ristein, J. & Ley, L. Recovery of surface conductivity of H-terminated diamond after thermal annealing in vacuum. *Phys. Rev. B* **69**, 125338 (2004).
- Bandis, C. & Pate, B. B. Photoelectric emission from negative-electron-affinity diamond (111) surfaces: Exciton breakup versus conduction-band emission. *Phys. Rev. B* **52**, 12056–12071 (1995).
- Golden, M. S. et al. The electronic structure of fullerenes and fullerene compounds from high energy spectroscopy. *J. Phys. Condens. Matter* **7**, 8219–8247 (1995).

Acknowledgements We acknowledge A. Hirsch for the gift of C_{60} and fruitful discussion. Partial financial support by the Deutsche Forschungsgemeinschaft is also gratefully acknowledged.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R. (juergen.ristein@physik.uni-erlangen.de).

The elasticity of the MgSiO_3 post-perovskite phase in the Earth's lowermost mantle

T. Iitaka¹, K. Hirose², K. Kawamura² & M. Murakami²

¹Computational Astrophysics Laboratory, RIKEN (The Institute of Physical and Chemical Research), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

²Department of Earth and Planetary Sciences, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152-8551, Japan

MgSiO_3 perovskite has been assumed to be the dominant component of the Earth's lower mantle, although this phase alone cannot explain the discontinuity in seismic velocities observed 200–300 km above the core–mantle boundary (the D'' discontinuity) or the polarization anisotropy observed in the lowermost mantle¹. Experimental and theoretical studies that have attempted to attribute these phenomena to a phase transition in the perovskite phase have tended to simply confirm the stability of the perovskite phase^{2–6}. However, recent *in situ* X-ray diffraction measurements have revealed⁷ a transition to a 'post-perovskite' phase above 125 GPa and 2,500 K—conditions close to those at the D'' discontinuity. Here we show the results of first-principles calculations of the structure, stability and elasticity of both phases at zero temperature. We find that the post-perovskite phase becomes the stable phase above 98 GPa, and may be responsible for the observed seismic discontinuity and anisotropy in the lowermost mantle. Although our ground-state calculations of the unit cell do not include the effects of temperature and minor elements, they do provide a consistent explanation for a number of properties of the D'' layer.

The first-principles calculations were performed with the initial models of the perovskite and post-perovskite phases, which are the orthorhombic unit cells containing four molecules with the space group $Pbnm$ and $Cmcm$, respectively (Fig. 1). Then the lattice constants and the positions of the atoms were optimized within the given symmetry to minimize the enthalpy (see Methods). This procedure was repeated by changing the external pressure from 80 GPa to 130 GPa in steps of 10 GPa. It was found that the post-perovskite phase becomes the stable phase above 98 GPa at $T = 0$ K by drawing a graph of the enthalpy difference $\Delta H = H(\text{PP}) - H(\text{Pv})$ as a function of pressure (Fig. 2). (Here we use PP to indicate a property of the post-perovskite phase, and Pv a property of the perovskite phase.)

Table 1 lists the parameters of the optimized structures of the post-perovskite and perovskite phases at 120 GPa, together with the experimental data⁷. The coordination numbers of Mg and Si in the post-perovskite phase are same as those in the perovskite phase—namely, eight for Mg and six for Si. The principal difference between the two structures is the linkage of SiO_6 octahedra. In the post-perovskite phase, they share edges, thereby forming chains, and the chains share the octahedral corners thereby forming flat sheets⁷, whereas in the perovskite phase, the octahedra share corners, thereby forming a three-dimensional network. The average Si–O distance of the post-perovskite phase (1.665 Å) is longer than that of the perovskite phase (1.650 Å); on the other hand, the average Mg–O distance of the post-perovskite phase (1.959 Å) is shorter than that of the perovskite phase (1.981 Å). The reduction of the Mg–O polyhedra and both the edge sharing and deformation of the SiO_6 octahedra cause a volume reduction of the post-perovskite phase from that of the perovskite phase. The small space for cations suggests that Mg, which is a very small cation, is much more comfortable in post-perovskite phase than other cations. Our first-principles calculations show that the post-perovskite phase is

about 1.4% denser than the perovskite phase throughout the pressure range, which is in agreement with the experimental value⁷ of 1.0–1.2% at 120 GPa and 300 K.

The post-perovskite phase of MgSiO_3 (Fig. 1b) is isostructural with UFeS_3 , UFeSe_3 , ThMnTe_3 , ThMgTe_3 , UMnSe_3 and ThMnSe_3 (refs 8–10). The crystal habit of ThMnTe_3 and ThMgTe_3 was reported⁹ to be platy (and needle-like). Although the crystallographic direction was not mentioned in the literature, the surface of the plate should be the (010) surface, which is parallel to the sheet structure, and the direction of the needle should be the [100] direction, which is the direction of the edge-shared octahedral chain. There is an important difference between the post-perovskite phase of MgSiO_3 and those of $\text{Th}^{(\text{IV})}\text{Mn}^{(\text{II})}\text{Te}_3$ and $\text{Th}^{(\text{IV})}\text{Mg}^{(\text{II})}\text{Te}_3$: the eight-coordinated Mg–O bonds are weaker than $\text{Th}^{(\text{IV})}\text{Te}^{(\text{II})}$ and the octahedral Si–O bonds are stronger than $\text{Mn}^{(\text{II})}\text{Te}^{(\text{II})}$ and $\text{Mg}^{(\text{II})}\text{Te}^{(\text{II})}$. These bonding features imply that the inter-chain bonds of the MgSiO_3 post-perovskite phase are much stronger than those of the ThMnTe_3 and ThMgTe_3 post-perovskite phase, and that the MgSiO_3 post-perovskite phase most probably has the platy crystal habit with (010) facets. The system of uranium transition-metal chalcogenides UMS_3 and UMSe_3 (where M is a transition metal) also shows an interesting relation to the perovskite–post-perovskite phase transition, because the structure of these compounds changes when M changes, instead of when the pressure changes: UMS_3 ($M = \text{V}, \text{Cr}, \text{Co}, \text{Ni}, \text{Ru}, \text{Rh}$) and UMSe_3 ($M = \text{V}, \text{Cr}, \text{Co}, \text{Ni}$) take the $Pbnm$ structure, and UMnSe_3 , UFeSe_3 and UFeS_3 take the $Cmcm$ structure. The mechanism behind this

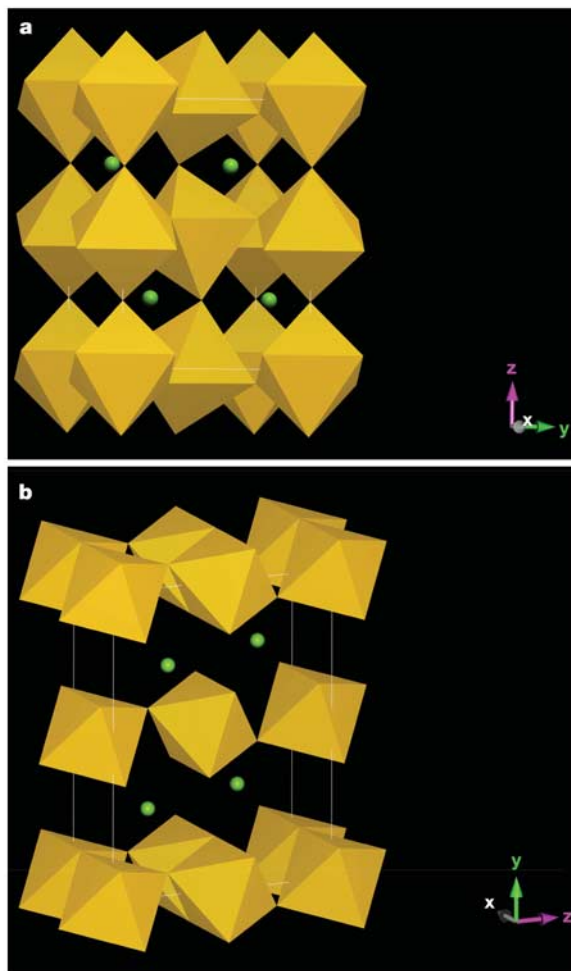


Figure 1 The unit cell structures of MgSiO_3 . **a**, Perovskite; **b**, post-perovskite. The spheres represent Mg, and the octahedra represent Si with sixfold oxygen coordination.

structure change is still unknown¹⁰.

Table 2 shows the nine elastic stiffness constants c_{ij} evaluated by the first-principles calculations. The crystal of the post-perovskite phase is much more compressible along the b -axis than along the a - or c -axis as $c_{22} \ll c_{11}, c_{33}$ (see also Supplementary Fig. 1), which could be a significant source of its seismic anisotropy. The propagation direction ($\mathbf{n}$) dependence of the seismic wave velocities in the perovskite and post-perovskite phase at the phase boundary (~ 100 GPa) were evaluated by solving the Christoffel equation $\det[c_{ijkl}n_jn_l - \rho V^2\delta_{ik}] = 0$ for the single crystal (Fig. 3). The wave velocities vary significantly with the propagation direction and the polarization, suggesting that both the phases exhibit strong anisotropy in both compressional (P) and shear (S) waves. The particular softness of the post-perovskite phase along the b -axis is reflected in the velocity minima when the polarization is along the b -axis, that is, when the propagation direction is along [010] for P waves or along [100] and [001] for S waves. A measure of anisotropy may be defined^{11,12} by $A = (v_{\max} - v_{\min})/\langle v \rangle \times 100\%$, where $v_{\max}$ and $v_{\min}$ are the maximum and minimum velocities, respectively, and $\langle v \rangle$ is the velocity for an isotropic aggregate of crystal calculated from the c_{ij} by the Voigt-Reuss-Hill averaging scheme¹. Though both the phases show strong anisotropy, the post-perovskite phase is found more anisotropic than the perovskite phase: the azimuth anisotropy of P and S waves is $A_P = 16\%$ and $A_S = 31\%$ for the post-perovskite phase, whereas $A_P = 11\%$ and $A_S = 16\%$ for the perovskite phase.

The D'' discontinuity is observed¹³ in many regions around the world about 200–300 km above the core–mantle boundary with an increase of up to 3.0% in both S- and P-wave velocities. On the basis of global seismic tomography, Sidorin¹⁴ proposed a model of a possibly ubiquitous seismic discontinuity atop the D'' layer that is caused by a solid–solid phase transition with a Clapeyron slope of 6 MPa K^{−1}. The increase of seismic velocities atop the D'' layer is attributed to the increase of bulk (K) and shear (G) moduli at the transition, which overcomes the increase of the density. One of the weak points of this model was that no relevant phase transition had been observed in the major constituents of the lower mantle, MgSiO₃ perovskite and magnesiowüstite. Therefore the discovery of the post-perovskite phase transition above 125 GPa and 2,500 K by Murakami *et al.*⁷ provides a strong ground for the model. The phase boundary drawn with the transition pressures at $T = 0$ K (from the first-principles calculation) and at $T = 2,500$ K (from the high-pressure experiments) has a Clapeyron slope of 10 MPa K^{−1}, which is close to Sidorin's value.

In the D'' layer significant polarization anisotropy has been reported^{15,16}, especially under the circum-Pacific regions, where the horizontally polarized S-wave velocity (v_{SH}) is 1–3% faster

than the vertically polarized S-wave velocity (v_{SV}). The post-perovskite phase transition can explain why the polarization anisotropy is observed only in the D'' layer but not in the overlying mantle¹⁵. In the possible horizontal shear flow in the D'' layer, crystals of the post-perovskite phase are probably preferentially oriented with the most compressible b -axis in the vertical direction. This seems very likely because the crystal of the post-perovskite phase has the sheet stacking structure along the b -axis (Fig. 1b). Therefore we modelled the phase boundary by using an isotropic aggregate of perovskite phase and a transversely isotropic aggregate of post-perovskite phase with the b -axis as the symmetry axis. Then we calculated the seismic wave velocities at the phase boundary by using the Voigt-Reuss-Hill averaging scheme for the perovskite phase, and the scheme of refs 17 and 18 for the velocities of the horizontally propagating P wave ($v_{PH}^{(PP)}$) and the horizontally and vertically polarized S-waves ($v_{SH}^{(PP)}, v_{SV}^{(PP)}$) in the post-perovskite phase. The discontinuities in P and S waves are estimated as 4–7% or $v_{PH}^{(PP)}/\langle v_P^{(Pv)} \rangle = 1.04$, $v_{SH}^{(PP)}/\langle v_S^{(Pv)} \rangle = 1.07$, and $v_{SV}^{(PP)}/\langle v_S^{(Pv)} \rangle = 1.03$. The polarization anisotropy of S waves in the post-perovskite phase is estimated as 4% or $v_{SH}^{(PP)}/v_{SV}^{(PP)} = 1.04$. This result may be regarded as consistent with the observations^{13,14} if we consider the effect of imperfect crystal alignments. With isotropic aggregates of the post-perovskite phase, very small discontinuities, $\langle v_P^{(PP)} \rangle / \langle v_P^{(Pv)} \rangle = 0.999$ and $\langle v_S^{(PP)} \rangle / \langle v_S^{(Pv)} \rangle = 1.01$, and no anisotropy are expected. Our model is in contrast to the model with the preferentially oriented perovskite phase where the most compressible axis becomes horizontal, and therefore the wrong polarization anisotropy, $v_{SV} > v_{SH}$, occurs^{18,19}.

Moreover, anti-correlation between bulk-sound velocity and S-wave velocity has been observed in the very deep portions of the mantle²⁰. The present results show slower bulk-sound velocity and

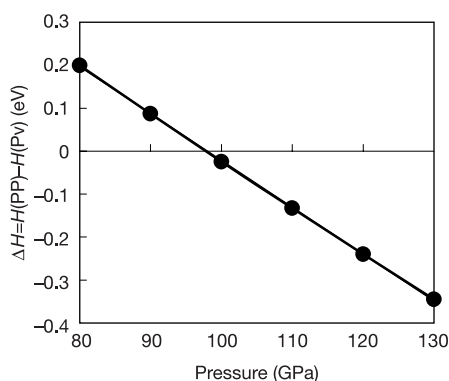


Figure 2 The enthalpy difference between the perovskite phase and post-perovskite phase as a function of pressure. The post-perovskite (PP) phase is favoured over the perovskite (Pv) phase at pressures above 98 GPa.

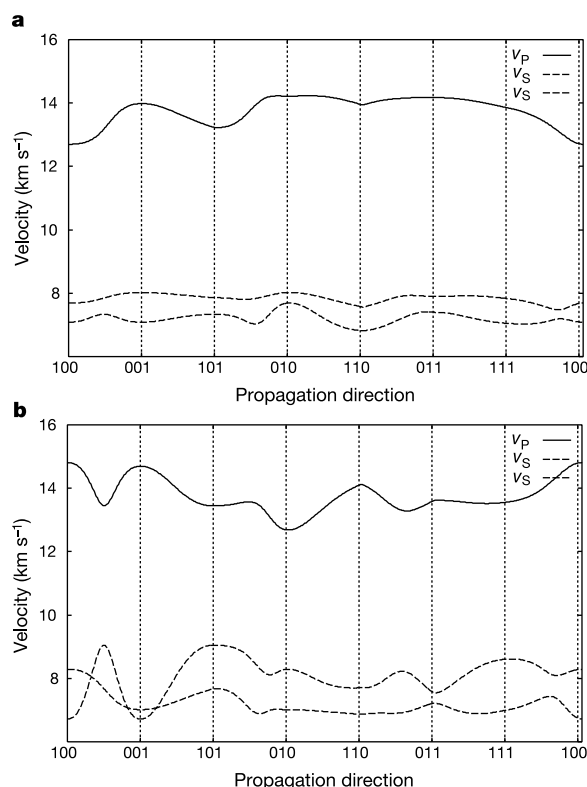


Figure 3 The variation of compressional (v_P) and shear (v_S) wave velocities as a function of propagation direction. **a**, Perovskite phase at 100 GPa; **b**, post-perovskite phase at 100 GPa. The two dashed lines represent the two polarizations of the shear waves.

Table 1 Crystal data of post-perovskite and perovskite phases

Coordinate		Post-perovskite phase		Perovskite phase
		Experimental ⁷ (121 GPa, 300 K)	Calculated (120 GPa, 0 K)	Calculated (120 GPa, 0 K)
Crystal system		Orthorhombic	Orthorhombic	Orthorhombic
Space group		<i>Cmcm</i> (no. 63)	<i>Cmcm</i>	<i>Pbnm</i> (no. 62)
Cell parameters				
<i>a</i> (Å)		2.456	2.455	4.289
<i>b</i> (Å)		8.042	8.051	4.557
<i>c</i> (Å)		6.093	6.099	6.264
<i>Z</i>		4	4	4
<i>V</i> (Å ³)		120.39	120.55	122.43
Atomic coordinates				
Mg (4c)	<i>x</i>	0.0	0.0	0.477
	<i>y</i>	0.253	0.253	0.925 (4c)
	<i>z</i>	0.25	0.25	0.25
Si (4a)	<i>x</i>	0.0	0.0	0.0
	<i>y</i>	0.0	0.0	0.0 (4a)
	<i>z</i>	0.0	0.0	0.0
O1 (4c)	<i>x</i>	0.0	0.0	0.884
	<i>y</i>	0.923	0.928	0.033 (4c)
	<i>z</i>	0.25	0.25	0.25
O2 (8f)	<i>x</i>	0.0	0.0	0.817
	<i>y</i>	0.631	0.636	0.307 (8d)
	<i>z</i>	0.436	0.442	0.557
Interatomic distances (Å)				
Si–O			1.632(×2), 0.681(×4)	1.645(×2), 1.649(×2) 1.657(×2)
Average			[1.665]	[1.650]
O–O edges			2.296, 2.305, 2.327, 2.337, 2.380, 2.456	2.313, 2.321, 2.330, 2.346, 2.347, 2.349, 2.415, 2.464
Mg–O			1.867(×2), 1.943(×4) 2.084(×2)	1.815, 1.848(×2), 1.885, 2.042(×2) 2.185(×2)
Average			[1.959]	[1.981]
Shortest Si–Si			2.455	3.129
Si–Mg			2.547	2.494
Mg–Mg			2.455	2.997

Table 2 The elastic-stiffness constants *c_{ij}* of post-perovskite and perovskite phases

	<i>C</i> ₁₁	<i>C</i> ₂₂	<i>C</i> ₃₃	<i>C</i> ₁₂	<i>C</i> ₁₃	<i>C</i> ₂₃	<i>C</i> ₄₄	<i>C</i> ₅₅	<i>C</i> ₆₆	<i>K</i>	<i>G</i>
Post-perovskite phase											
100 GPa	1175	862	1157	366	289	437	264	242	368	595	306
120 GPa	1270	937	1264	425	329	493	291	264	412	660	332
Perovskite phase											
100 GPa	852	1068	1033	461	368	391	340	265	312	596	296
120 GPa	909	1160	1128	520	410	434	363	278	339	654	314
Perovskite phase ¹⁸											
100 GPa	926	1160	1056	460	380	406	360	265	330	623	314
120 GPa	996	1263	1156	530	415	449	388	279	360	682	337

The values of Wentzcovitch¹⁸ are interpolated to third-order polynomials by using the data at five different pressures listed in their paper.

faster S-wave velocity for the post-perovskite phase than for the perovskite phase at equivalent pressure, suggesting that such negative correlation may be accounted for by the heterogeneity in the mineral proportion of the perovskite phase and the post-perovskite phase. It is, however, noted that the present calculations were made at *T* = 0 K and for the Mg end-member composition. The effects of temperature^{21,22} and the solution of minor elements were neglected here. The experiments on natural pyrolitic mantle composition showed that the MgSiO₃-rich post-perovskite contains much less iron than the perovskite²³, in accordance with our ‘space for cations’ argument. □

Methods

Our first-principles calculations were performed with CASTEP 4.2 codes based on the plane wave basis set, the Vanderbilt-type ultrasoft pseudopotentials^{24–26} for electron–ion interaction, and the local density approximation (LDA) for exchange–correlation interaction. Previous studies^{11,12,18,26} show that such a method reproduces very well the structures and properties of silicates under high pressure. The Brillouin zones are sampled with 6 × 4 × 4 Monkhorst-Pack k-points²⁷ for the perovskite phase and 8 × 2 × 4 k-points for the post-perovskite phase by using the maximal symmetry of the model. Increasing the number of k-points to 6 × 6 × 4 and 10 × 4 × 4, respectively, changes the total energy only 10^{−5} eV per atom. The cut-off energy is set to 800 eV; increasing the cut-off energy to 1,600 eV changes the enthalpy difference of the two phases by only 10^{−3} eV per atom. During the structural optimization, the enthalpy *H* = *E* + *PV* is minimized by varying cell shape and atomic positions under the restriction of the given symmetry. In the geometrical optimization, the total stress tensor²⁸ is reduced to the order of 0.01 GPa by

using the finite basis-set corrections²⁹. The nine elastic stiffness constants *c_{ij}* are evaluated by computing the stress tensor *σ_i* generated by forcing strain *ε_j* to the optimized unit cell with four different magnitudes, *ε* = −0.02, −0.01, 0.01 and 0.02, and by fitting them to the third-order polynomial of *ε* to remove the nonlinear contributions^{11,12,30}.

Received 10 March; accepted 27 May 2004; doi:10.1038/nature02702.

1. Poirier, J. P. *Introduction to the Physics of the Earth's Interior* (Cambridge Univ. Press, Cambridge, 2000).

2. Fiquet, G., Dewaele, A., Andrault, D., Kunz, M. & Le Bihan, T. Thermoelastic properties and crystal structure of MgSiO₃ perovskite at lower mantle pressure and temperature conditions. *Geophys. Res. Lett.* **27**, 21–24 (2000).

3. Andrault, D. Evaluation of (Mg, Fe) partitioning between silicate perovskite and magnesiowüstite up to 120 GPa and 2300 K. *J. Geophys. Res.* **106**, 2079–2087 (2001).

4. Stixrude, L. & Cohen, R. E. Stability of orthorhombic MgSiO₃ perovskite in the Earth's lower mantle. *Nature* **364**, 613–616 (1993).

5. Wentzcovitch, R. M., Ross, N. L. & Price, G. D. Ab initio study of MgSiO₃ and CaSiO₃ perovskites at lower-mantle pressures. *Phys. Earth Planet. Inter.* **90**, 101–112 (1995).

6. Warren, M. C., Ackland, G. J., Karki, B. B. & Clark, S. J. Phase transitions in silicate perovskites from first principles. *Mineral. Mag.* **62**, 585–598 (1998).

7. Murakami, M., Hirose, K., Kawamura, K., Sata, N. & Ohishi, Y. Post-perovskite phase transition in MgSiO₃. *Science* **304**, 855–858 (2004).

8. Noel, H. & Padiou, J. Structure crystalline de FeUS₃. *Acta Crystallogr. B* **32**, 1593–1595 (1976).

9. Narducci, A. A. & Ibers, J. A. The related compounds MThTe₃ (M = Mn, Mg) and ACuThSe₃ (A = K, Cs): Syntheses and characterization. *Inorg. Chem.* **39**, 688–691 (2000).

10. Ijjaali, I., Mitchell, K., Huang, F. Q. & Ibers, J. A. Syntheses and characterization of the actinide manganese selenides ThMnSe₃ and U₂MnSe₃. *J. Solid State Chem.* **177**, 257–261 (2004).

11. Karki, B. B. *et al.* Structure and elasticity of MgO at high pressure. *Am. Mineral.* **82**, 51–60 (1997).

12. Karki, B. B. *et al.* Elastic properties of orthorhombic MgSiO₃ perovskite at lower mantle pressures. *Am. Mineral.* **82**, 635–638 (1997).

13. Wyssession, M. E., *et al.* in *The Core-Mantle Boundary Region* (eds Gurnis, M., Wyssession, M. E., Knittle, E. & Buffett, B. A.) 273–297 (American Geophysical Union, Washington DC, 1998).
14. Sidorin, I., Gurnis, M. & Helmberger, D. V. Evidence for a ubiquitous seismic discontinuity at the base of the mantle. *Science* **286**, 1326–1331 (1999).
15. Lay, T., Williams, Q., Garnero, E. J., Kellogg, L. & Wyssession, M. E. in *The Core-Mantle Boundary Region* (eds Gurnis, M., Wyssession, M. E., Knittle, E. & Buffett, B. A.) 299–318 (American Geophysical Union, Washington DC, 1998).
16. Panning, M. & Romanowicz, B. Inferences on flow at the base of Earth's mantle based on seismic anisotropy. *Science* **303**, 351–353 (2004).
17. Montagner, J. P. & Nataf, H. C. A simple method for inverting the azimuthal anisotropy of surface waves. *J. Geophys. Res.* **91**, 511–520 (1986).
18. Wentzcovitch, R. M., Karki, B. B., Karato, S. & Da Silva, C. R. S. High pressure elastic anisotropy of MgSiO₃ perovskite and geophysical implications. *Earth Planet. Sci. Lett.* **164**, 371–378 (1998).
19. Karato, S., Zhang, S. & Wenk, H. R. Superplasticity in Earth's lower mantle: evidence from seismic anisotropy and rock physics. *Science* **270**, 458–461 (1995).
20. Masters, G. & Laske, G. in *Earth's Deep Interior: Mineral Physics and Tomography from the Atomic to the Global Scale* (eds Karato, S., Forte, A., Liebermann, R., Masters, G. & Stixrude, L.) 63–87 (American Geophysical Union, Washington DC, 2000).
21. Oganov, A. R., Brodholt, J. P. & Price, G. D. The elastic constants of MgSiO₃ perovskite at pressures and temperatures of the Earth's mantle. *Nature* **411**, 934–937 (2001).
22. Wentzcovitch, R. M., Karki, B. B., Cococcioni, M. & de Gironcoli, S. Thermoelastic properties of MgSiO₃-perovskite: insights on the nature of the Earth's lower mantle. *Phys. Rev. Lett.* **92**, 018501 (2004).
23. Murakami, M. *Phase Transition of Lower Mantle Mineral and its Geophysical Implications* Thesis, Tokyo Institute of Technology (2004).
24. Vanderbilt, D. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B* **41**, 7892–7895 (1990).
25. Civalleri, B. & Harrison, N. M. New ultrasoft pseudopotentials for the study of silicates. *Mol. Simulat.* **28**, 213–237 (2002).
26. Brodholt, J. P., Oganov, A. R. & Price, G. D. Computational mineral physics and the physical properties of perovskite. *Phil. Trans. R. Soc. Lond. A* **360**, 2507–2520 (2002).
27. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
28. Nielsen, O. H. & Martin, R. M. First principles calculation of stress. *Phys. Rev. Lett.* **50**, 697–700 (1983).
29. Francis, G. P. & Payne, M. C. Finite basis set corrections to total energy pseudopotential calculations. *J. Phys. Condens. Matter* **2**, 4395–4404 (1990).
30. Iitaka, T. & Ebisuzaki, T. First-principles calculation of elastic properties of solid argon at high pressures. *Phys. Rev. B* **65**, 012103 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Kaneshima for discussions, D.M. Bird for providing CASTEP codes, N.M. Harrison for pseudopotentials and the computer centres of RIKEN and NIG for access to the supercomputers. This work was also supported by JASRI/Spring-8 and IFREE/JAMSTEC.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.I. (tiitaka@riken.jp).

Theoretical and experimental evidence for a post-perovskite phase of MgSiO₃ in Earth's D'' layer

Artem R. Oganov¹ & Shigeaki Ono²

¹Laboratory of Crystallography, Department of Materials, ETH Zurich, Wolfgang Pauli Strasse 10, CH-8093 Zurich, Switzerland

²Institute for Frontier Research on Earth Evolution, Japan Agency for Marine-Earth Science and Technology, 2-15 Natsushima-cho, Yokosuka-shi, Kanagawa 237-0061, Japan

The Earth's lower mantle is believed to be composed mainly of (Mg,Fe)SiO₃ perovskite, with lesser amounts of (Mg,Fe)O and CaSiO₃ (ref. 1). But it has not been possible to explain many unusual properties of the lowermost ~150 km of the mantle (the D'' layer) with this mineralogy. Here, using *ab initio* simulations and high-pressure experiments, we show that at pressures and

temperatures of the D'' layer, MgSiO₃ transforms from perovskite into a layered CaIrO₃-type post-perovskite phase. The elastic properties of the post-perovskite phase and its stability field explain several observed puzzling properties of the D'' layer: its seismic anisotropy², the strongly undulating shear-wave discontinuity at its top^{3–6} and possibly the anticorrelation between shear and bulk sound velocities^{7,8}.

If MgSiO₃ perovskite is stable throughout the lower mantle, it should be the most abundant mineral in our planet. While some researchers⁹ have suggested its decomposition into the oxides at lower mantle conditions, most workers^{10–13} have found that perovskite is more stable than the oxides. To our knowledge, the possibility that MgSiO₃ could be stable in a completely new structure within the lower mantle has not been considered.

The shear-wave discontinuity at the top of the D'' layer, suggested in ref. 3, has a strong topography. This discontinuity has commonly been explained by some chemical difference between the D'' layer and the rest of the lower mantle. However, using a combination of dynamical and seismic modelling, Sidorin *et al.*^{4–6} have shown that the most consistent explanation is a phase transition in mantle minerals. Their best model^{5,6} had a shear-wave discontinuity of ~1% located ~150 km above the core-mantle boundary (depth 2,740 km), with a Clapeyron slope of 6 MPa K⁻¹ (though values as large as 10 MPa K⁻¹ were almost equally acceptable). The discontinuities of the compressional wave velocities and of the density could not be resolved. The models of Sidorin *et al.*^{4–6} were very appealing, but the major problem was that no appropriate phase transition was known at the time. Here we show that MgSiO₃ perovskite undergoes a structural phase transition at the conditions corresponding to the top of the D'' layer. The predicted seismic signatures of this transition match the seismological inferences of Sidorin *et al.*^{4–6}.

A key observation made by Ono *et al.*¹⁴ was that Fe₂O₃, like MgSiO₃, transforms from the corundum (or ilmenite) to the perovskite structure under pressure. As these authors further found¹⁴, a post-perovskite phase of Fe₂O₃ with a CaIrO₃-type *Cmcm* structure¹⁵ (Fig. 1) is stable above 60 GPa. This has led to the idea that a similar structure could be stable for MgSiO₃ at high pressure.

We explored this idea using *ab initio* simulations based on density functional theory within the local density approximation

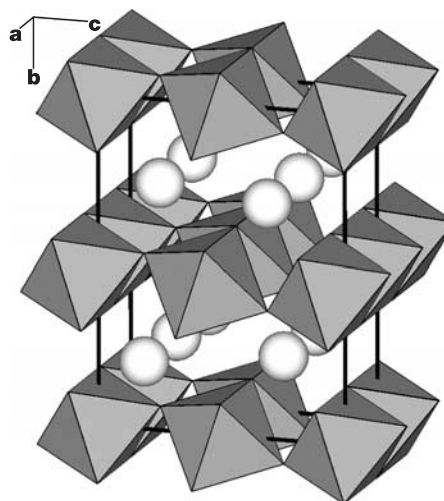


Figure 1 Structure of the post-perovskite phase of MgSiO₃ (calculated at 120 GPa). SiO₆ octahedra and Mg atoms (spheres) are shown. Similar structures are known for Fe₂O₃, CaIrO₃, FeUS₃, PbTiI₃, US₂Si₃, KTlI₃, AgTaS₃ and CaInBr₃.

strain relations at 120 GPa are given in Table 3. The density discontinuity at the transition is 1.4%; for the mantle the expected density discontinuity is 1.1%. The predicted shear wave discontinuity in MgSiO_3 is 1.9% (1.4% for the mantle), consistent with $\sim 1\%$ suggested by Sidorin *et al.*^{4–6}. We predict a very small discontinuity for the compressional wave velocity (0.3% for pure MgSiO_3), explaining why it was so seldom found at the top of the D'' layer^{4–6}.

The calculated properties of our new phase explain many puzzling features of the D'' layer. In seismological models, horizontally polarized shear waves are faster (by 1% on average²) than the vertically polarized ones ($v_{\text{SH}} > v_{\text{SV}}$). Significant seismic anisotropy of the D'' layer, containing a signature of the convective flow², could not be explained even qualitatively by mineral physics^{23,24}. The structure of the post-perovskite phase has silicate layers parallel to {010}, which is then the most natural slip plane that will be oriented parallel to the convective flow. In regions of horizontal flow (most of D''), we obtain $v_{\text{SH}}/v_{\text{SV}} = 1.029 > 1$ for post-perovskite (using the method of ref. 25 and data of Table 3). This is consistent with seismological evidence, and suggests either a large degree of lattice-preferred orientation or significant contributions from other sources such as shape-preferred orientation (anisotropy due to the ordered distribution of crystals and inclusions in the rock). In regions of upwelling convective streams (for example, below the central Pacific), slip planes would be predominantly vertical and one would obtain $v_{\text{SH}} < v_{\text{SV}}$ actually observed in such regions².

Another mystery of D'', the anticorrelation between the shear (v_{S}) and bulk sound (v_{P}) velocities^{7,8}, can also be quantitatively explained by the phase transition of MgSiO_3 from perovskite into the post-perovskite phase. We recall that it has been difficult to explain why, at a given depth in the lowermost mantle, anomalies of v_{S} and v_{P} have opposite signs. As Table 3 shows, the post-perovskite

transition has a positive jump of v_{S} and a negative jump of v_{P} . Note that the transition is first order, and in a multicomponent system such as the Earth's mantle there will be an interval ΔT of coexistence of the two phases at given pressure. For this two-phase region, we can write:

$$\left(\frac{\partial v_{\phi}}{\partial v_{\text{S}}}\right)_P \approx \left(\frac{\partial v_{\phi}}{\partial T}\right)_{P,x} + f \frac{v_{\phi 2} - v_{\phi 1}}{\Delta T} \quad (1)$$

which includes purely thermal responses (at given composition x and pressure P) of the velocities and effects due to a phase transition; f is the volume fraction of MgSiO_3 (75%). The poorly known effects of variation of Al and Fe content are not included in equation (1), but we find that anticorrelation can be explained without them. Using equation (1), data of Table 3, and assuming thermal responses equal to those of perovskite²⁰, we get $(\partial \ln v_{\text{S}} / \partial \ln v_{\text{P}})_P = -0.15$ and -0.33 for $\Delta T = 250$ K and 50 K, respectively (from seismic tomography: -0.15 (ref. 8) and -0.3 (ref. 7)). Furthermore, we can reproduce the positive correlation between the shear and compressional (v_{P}) velocities with the same ΔT : $(\partial \ln v_{\text{S}} / \partial \ln v_{\text{P}})_P = 3.36$ for $\Delta T = 250$ K (3.3 from ref. 8).

The fact that so many previously unexplained seismic features of the D'' layer (seismic discontinuity, its magnitude and Clapeyron slope, anisotropy, bulk-shear velocity anticorrelation) are naturally explained by the properties of post-perovskite is a strong indication that this phase is indeed the major component of the D'' layer. The D'' layer is not necessarily chemically very different from the rest of the lower mantle, but it surely is different mineralogically. Neither theoretical and experimental error bars (a few GPa for transition pressure) nor the effects of temperature as explored here would change our prediction that the CaIrO_3 -type post-perovskite phase is stable in the D'' layer. At present, very little is known about the effects of Fe^{2+} , Fe^{3+} and Al^{3+} impurities on mantle minerals. We expect that at least Fe^{3+} impurities should stabilize the post-perovskite phase against perovskite; this is because at high pressure Fe_2O_3 also has the CaIrO_3 structure¹⁴.

Our findings have other implications. For instance, rheological properties of post-perovskite (probably very different from those of perovskite) and the predicted density discontinuity (1.1%) at top of the D'' layer could be important for mantle dynamics. Also, element partitioning between post-perovskite, perovskite, and molten Fe might be a key to some geochemical anomalies. Elements that are incompatible in the mantle (for example, Na, K, U, Th) might be more easily accommodated in the layered post-perovskite structure, which may affect the chemistry of plume magmas. As Fe_2O_3 at high pressure has the same structure as post-perovskite, post-perovskite could have a greater concentration of Fe^{3+} than perovskite. It is also likely that the size of the D'' region increased significantly with time as the mantle cooled down; this is because the Clapeyron slope of the post-perovskite transition is large, $8.0\text{--}9.6 \text{ MPa K}^{-1}$. If the whole mantle had been molten with temperatures above $4,000\text{--}4,500$ K in the early history of the Earth, it would be perovskite that crystallized from the cooling melt, and only on further cooling would post-perovskite and the D'' layer have appeared. The present-day thickness of the D'' could be used to estimate its age, given the cooling history of the lowermost mantle. As post-perovskite stability requires pressures unattainable in smaller planets like Mercury

Table 1 Structures of post-perovskite and perovskite at 120 GPa

Post-perovskite*				Perovskite†		
Mg	0	0.2532	1/4	Mg	0.5246	0.5768
Si	0	0	0	Si	1/4	0
O1	0	0.9276	1/4	O1	0.1164	0.4669
O2	0	0.6356	0.4413	O2	0.1829	0.1926

Table shows GGA results.

*Space group $Cmcm$: $a = 2.474 \text{ \AA}$, $b = 8.121 \text{ \AA}$, $c = 6.138 \text{ \AA}$; distances (in $\text{\AA}$): $\text{Mg-O1} = 1.880$ ($\times 2$), $\text{Mg-O2} = 1.955$ ($\times 4$), 2.099 ($\times 2$); $\text{Si-O1} = 1.643$ ($\times 2$), $\text{Si-O2} = 1.695$ ($\times 4$).

†Space group $Pbnm$: $a = 4.318 \text{ \AA}$, $b = 4.595 \text{ \AA}$, $c = 6.305 \text{ \AA}$; distances (in $\text{\AA}$): $\text{Mg-O1} = 1.833$ ($\times 1$), 1.893 ($\times 1$), $\text{Mg-O2} = 1.864$ ($\times 2$), 2.047 ($\times 2$), 2.201 ($\times 2$); $\text{Si-O1} = 1.661$ ($\times 2$), $\text{Si-O2} = 1.659$ ($\times 2$), 1.670 ($\times 2$).

Table 2 Vinet equations of state of perovskite and post-perovskite

Parameters	Post-perovskite		Perovskite		Exp.*
	LDA	GGA	LDA	GGA	
V_0 ($\text{\AA}^3$)	162.86	167.64	163.35	167.42	162.3
K_0 (GPa)	231.93	199.96	259.82	230.05	259.5
K_0'	4.430	4.541	4.060	4.142	3.69

*Third-order Birch-Murnaghan equation of state parameters at room temperature¹⁰.

Table 3 Elastic constants of perovskite and post-perovskite at 120 GPa

	C_{11}	C_{22}	C_{33}	C_{12}	C_{13}	C_{23}	C_{44}	C_{55}	C_{66}	K	G
Perovskite*	907	1,157	1,104	513	406	431	364	271	333	648.0	310.9
Post-perovskite†	1,252	929	1,233	414	325	478	277	266	408	647.2	327.5

Table shows GGA results. All elastic constants are in GPa.

*Acoustic velocities (m s^{-1}): $v_{\text{P}} = 14,118$; $v_{\text{S}} = 7,636$; $v_{\text{P}} = 11,026$.

†Acoustic velocities (m s^{-1}): $v_{\text{P}} = 14,158$; $v_{\text{S}} = 7,783$; $v_{\text{P}} = 10,940$.

and Mars, many features of these planets would be different from those of the Earth. Further studies are necessary to address these and other issues (elasticity and anelasticity, electrical conductivity, radiative conductivity, energetics of stacking faults, effects of impurities on stability and properties of post-perovskite). Finally, we note that the results of a recent, independent, experimental study²⁶ of the post-perovskite phase transition are consistent with our theoretical and experimental findings. □

Received 24 March; accepted 27 May 2004; doi:10.1038/nature02701.

1. Fiquet, G. Mineral phases of the Earth's mantle. *Z. Krist.* **216**, 248–271 (2001).
2. Panning, M. & Romanowicz, B. Inferences on flow at the base of Earth's mantle based on seismic anisotropy. *Science* **303**, 351–353 (2004).
3. Lay, T. & Helmberger, D. V. A shear velocity discontinuity in the lower mantle. *Geophys. Res. Lett.* **10**, 63–66 (1983).
4. Sidorin, I., Gurnis, M., Helmberger, D. V. & Ding, X. Interpreting D' seismic structure using synthetic waveforms computed from dynamic models. *Earth Planet. Sci. Lett.* **163**, 31–41 (1998).
5. Sidorin, I., Gurnis, M. & Helmberger, D. V. Evidence for a ubiquitous seismic discontinuity at the base of the mantle. *Science* **286**, 1326–1331 (1999).
6. Sidorin, I., Gurnis, M. & Helmberger, D. V. Dynamics of a phase change at the base of the mantle consistent with seismological observations. *J. Geophys. Res.* **104**, 15005–15023 (1999).
7. Su, W. J. & Dziewonski, A. M. Simultaneous inversion for 3-D variations in shear and bulk velocity in the mantle. *Phys. Earth Planet. Inter.* **100**, 135–156 (1997).
8. Masters, G. *et al.* in *Earth's Deep Interior: Mineral Physics and Tomography from the Atomic to the Global Scale* (ed. Karato, S.-i.) 63–87 (AGU Geophysical Monograph 117, American Geophysical Union, Washington DC, 2000).
9. Saxena, S. K. *et al.* Stability of perovskite (MgSiO₃) in the Earth's mantle. *Science* **274**, 1357–1359 (1996).
10. Fiquet, G., Dewaele, A., Andrault, D., Kunz, M. & Le Bihan, T. Thermoelastic properties and crystal structure of MgSiO₃ perovskite at lower mantle pressure and temperature conditions. *Geophys. Res. Lett.* **27**, 21–24 (2000).
11. Serghiou, G., Zerr, A. & Boehler, R. (Mg,Fe)SiO₃-perovskite stability under lower mantle conditions. *Science* **280**, 2093–2095 (1998).
12. Shim, S. H., Duffy, T. S. & Shen, G. Y. Stability and structure of MgSiO₃ perovskite to 2300-kilometer depth in Earth's mantle. *Science* **293**, 2437–2440 (2001).
13. Ono, S., Ohishi, Y. & Mibe, K. Phase transition of Ca-perovskite and stability of Al-bearing Mg-perovskite in the lower mantle. *Am. Mineral.* (in the press).
14. Ono, S., Sata, N. & Ohishi, Y. Phase transformation of perovskite structure in Fe₂O₃ at high pressures and high temperatures. *Am. Mineral.* (submitted).
15. Rodi, F. & Babel, D. Erdalkaliridium(IV) - oxide: Kristallstruktur von CaIrO₃. *Z. Anorg. Allg. Chem.* **336**, 17–23 (1965).
16. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
17. Oganov, A. R. & Brodholt, J. P. High-pressure phases in the Al₂SiO₅ system and the problem of Al-phase in Earth's lower mantle: ab initio pseudopotential calculations. *Phys. Chem. Miner.* **27**, 430–439 (2000).
18. Baroni, S., de Gironcoli, S., Dal Corso, A. & Gianozzi, P. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **73**, 515–562 (2001).
19. Urusov, V. S. *Theoretical Crystal Chemistry* (Moscow State Univ. Press, Moscow, 1987) [in Russian].
20. Oganov, A. R., Brodholt, J. P. & Price, G. D. The elastic constants of MgSiO₃ perovskite at pressures and temperatures of the Earth's mantle. *Nature* **411**, 934–937 (2001).
21. Oganov, A. R., Brodholt, J. P. & Price, G. D. Ab initio elasticity and thermal equation of state of MgSiO₃ perovskite. *Earth Planet. Sci. Lett.* **184**, 555–560 (2001).
22. Ono, S., Hirose, K., Ishiki, M., Mibe, K. & Saito, Y. Equation of state of hexagonal aluminous phase of natural composition to 63 GPa at 300 K. *Phys. Chem. Miner.* **29**, 527–531 (2002).
23. Jeanloz, R. & Williams, Q. The core-mantle boundary region. *Rev. Mineral.* **37**, 241–259 (1998).
24. Wentzovitch, R. M., Karki, B. B., Karato, S. & da Silva, C. R. S. High pressure elastic anisotropy of MgSiO₃ perovskite and geophysical implications. *Earth Planet. Sci. Lett.* **164**, 371–378 (1998).
25. Montagner, J.-P. & Nataf, H.-C. A simple method for inverting the azimuthal anisotropy of surface waves. *J. Geophys. Res.* **91**, 511–520 (1986).
26. Murakami, M., Hirose, K., Kawamura, K., Sata, N. & Ohishi, Y. Post-perovskite phase transition in MgSiO₃. *Science* **304**, 855–858 (2004).
27. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
28. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
29. Kresse, G. & Furthmüller, J. Efficiency of ab initio total-energy calculations for metals and semiconductors using a plane-wave basis set. *Comp. Mater. Sci.* **6**, 15–50 (1996).
30. Gonze, X. *et al.* First-principles computation of materials properties: the ABINIT software project. *Comp. Mater. Sci.* **25**, 478–492 (2002).

Acknowledgements Calculations were performed at CSCS (Manno) and ETH Zurich. We thank P. Ulmer, A.N. Halliday, S. Goes, F. Cammarano, A.B. Thompson and P.J. Tackley for discussions, and Y. Ohishi and N. Sata for experimental support. Synchrotron radiation experiments were performed at the BL10XU, SPRING-8.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.R.O. (a.oganov@mat.ethz.ch).

Audience drives male songbird response to partner's voice

Clémentine Vignal^{1,2}, Nicolas Mathevon¹ & Stéphane Mottin²

¹Equipe 'Communications Acoustiques' NAMC CNRS UMR8620, Université Paris XI-Orsay and LBA Université Jean Monnet, 42023 Saint-Etienne Cedex 2, France

²TSI CNRS UMR5516 Université Jean Monnet, 42023 Saint-Etienne Cedex 2, France

According to the social intelligence hypothesis, social context represents an important force driving the selection of animal cognitive abilities such as the capacity to estimate the nature of the social relationships between other individuals^{1–4}. Despite this importance, the influence of this force has been assessed only in primates and never in other animals showing social interactions^{5–7}. In this way, avian communication generally takes place in a network of signallers and receivers, which represents an audience altering individual signalling behaviours^{8,9}. Indeed, vocal amplitude¹⁰ and repertoire¹¹ are known to be socially regulated and the attitude towards the opposite sex may change depending on the audience^{8,12,13}. This 'audience effect'^{8,14–16} provides support for the reality of social awareness in some bird species. However no evidence has yet been found to suggest that birds are able to estimate the characteristics of the social relationships between group-mates. Here we show that the male of a gregarious songbird species—the zebra finch (*Taeniopygia guttata*)—pays attention to the mating status of conspecific pairs, and uses this information to control its behaviour towards its female partner.

Zebra finches are monogamous flock-forming birds that seem to use acoustic recognition for pair-bond maintenance^{11,17}. A number of different vocalizations are produced by this species¹¹, distance calls being the most frequently emitted by both males and females. Distance calls are used by the members of a pair to remain in contact when the flock is foraging or feeding and especially when the birds lose visual contact with each other¹¹. As in many gregarious species, vocal recognition is thus likely to be a key component of reproductive success and it should be supported in both sexes by acoustic cues of distance calls^{18–20}. Previous laboratory experiments testing isolated birds demonstrated that the female zebra finch is able to recognize its mate's vocalizations from other males' vocalizations¹⁷, but never succeeded in demonstrating a mutual acoustic recognition between mates¹¹. In the natural biological context described above, it is very unlikely that wild male zebra finches do not recognize their mates' voices. Two main hypotheses can thus be envisaged: either captive zebra finches have lost some cognitive capacities because of domestication (for instance, females' calls may be less individualized, and/or males may no longer be able to recognize them), or tested males do not show preferential response to their mate's voice owing to a modification of their natural behaviour by their socially isolated position during the playback tests. Indeed social isolation could be a situation of stress in comparison to the natural context where the zebra finch lives in large groups and experiences permanent social interactions that may influence mate-directed behaviour.

To determine whether the vocalizations of female zebra finches support mate recognition, we analysed the acoustic structure of distance calls, searching for acoustic cues which could encode the emitter's individual identity. The female distance call is a complex sound with a fundamental frequency associated with several harmonics (Fig. 1a). This sound is frequency- and amplitude-modulated. With reference to frequency-modulation characteristics, the distance call can be divided into three segments of

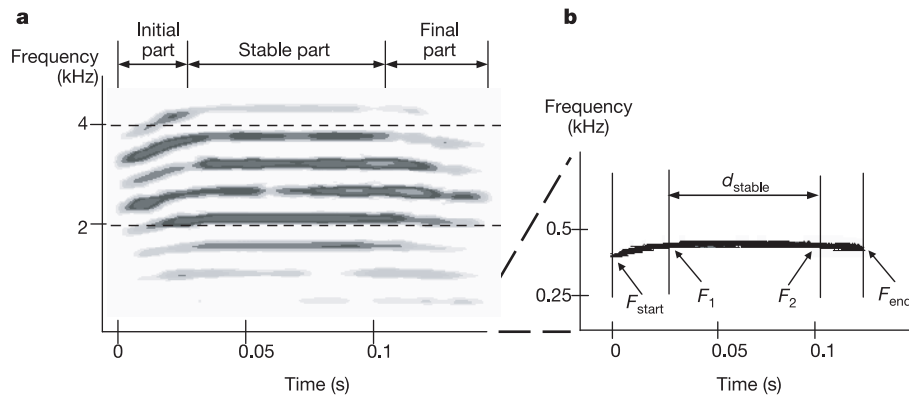


Figure 1 Acoustic structure of the distance call of the female zebra finch. **a**, Spectrogram of the call of a female zebra finch. The call can be divided into an initial part (rapid and loud ascending frequency modulation), a long and loud second part with no frequency modulation (the stable part) and a final part (rapid and soft descending frequency

modulation). The amplitude modulations are represented by a greyscale. **b**, Acoustic cues characterizing the fundamental frequency: the beginning frequency (F_{start}), the frequencies at the start and at the end of the second part (F_1 , F_2), the frequency at the end of the third part (F_{end}), and the duration of the second part (d_{stable}).

different durations (Fig. 1a): the initial segment defined by a rapid and loud ascending frequency modulation, a long and loud second segment with no frequency modulation (the stable part), and a third segment defined by a rapid and soft descending frequency modulation. To assess which of these spectral, amplitude and temporal cues could support individual identity coding and thus vocal recognition by males, we defined a set of 17 parameters describing the acoustic structure of the call, and measured the intra-individual and inter-individual variability of each parameter with a non-parametric analysis of variance. For each measured parameter the difference between the females is significant (Kruskal–Wallis ANOVA, $P < 0.05$) and the variation within individuals is smaller than that among individuals. In particular, the four spectral parameters (F_{start} , F_1 , F_2 , F_{end} ; Fig. 1b) describing the frequency modulation of the fundamental frequency are highly individualized (potential for individual identity coding values (PIC) > 2 ; see

Methods). The discriminant analysis based on the whole set of analysed acoustic parameters allows us to discriminate among 100% of the seven females (Fig. 2). According to our analysis, it thus appears that female distance calls contain potential cues to support individual identity coding and thus recognition by males.

To investigate whether mate recognition by males is affected by the current social context, we tested to see whether the vocal response of males to their partner's calls and to the calls of a familiar female changes depending on the composition of the audience. We performed playback experiments with 15 male zebra finches. Each of these birds was paired for several months with a different female and raised at least one brood. The tested male was separated from its mate, put in an experimental cage one night before the test, and assigned to one of the following social contexts: (1) the 'unmated males' context: two single males were placed in a companion cage near the experimental cage, (2) the 'mated pair' context: a normal male–female pair was placed in the companion cage, and (3) the 'unmated pair' context: a male and a female that were not paired were placed in two different companion cages. During playback, the

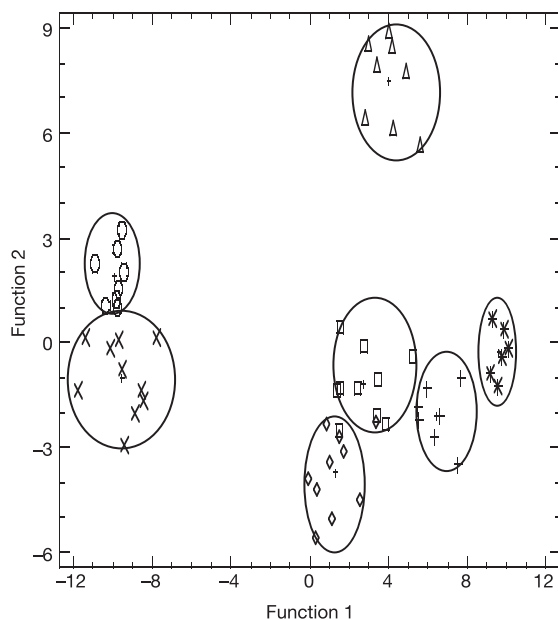


Figure 2 Discriminant analysis on the 17 acoustic parameters of the female call. The calls of seven individuals are represented according to the two main functions of the analysis. Each female is represented by one symbol and each point corresponds to one analysed call. The seven females are discriminated, so the distance calls could support individual identity coding.

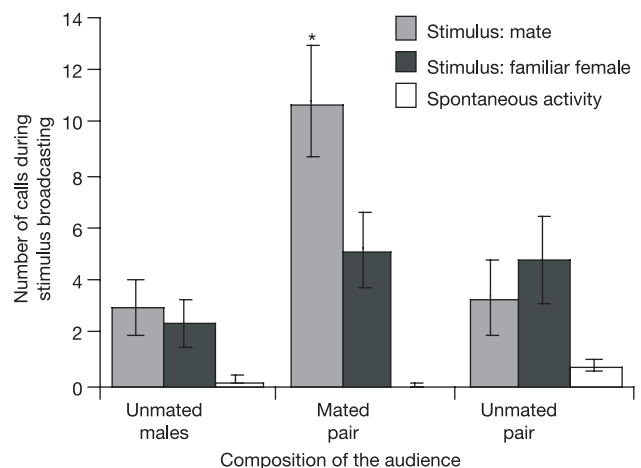


Figure 3 Response of male to female voices as a function of the mating status of the audience. The tested males were placed in different social contexts: (1) 'unmated males' (two single males placed in a companion cage), (2) 'mated pair' (a normal male–female pair in a companion cage), and (3) 'unmated pair' (a male and a female, not paired, in two different companion cages). Asterisk, significant difference between the male responses to mate and familiar female in the 'mated pair' context. The 'spontaneous' activity of the male (in the absence of any playback stimulus) is reported. Error bars indicate standard errors.

male, accompanied by its audience, was presented with two sets of stimuli: a series of calls of its mate and a series of calls of a familiar female. The male's response was assessed by counting the number of calls emitted during the stimulus broadcasting. The audience behaviour in response to stimuli was also monitored. Before the playback, the 'spontaneous' activity of the male was always very low (Fig. 3) and not dependent on the audience (Kruskall-Wallis ANOVA, $P = 0.069$). Whereas the audience behaviour seems to be fully independent of which female's calls were being presented (Wilcoxon matched-pairs test: 'unmated males', $P = 0.892$; 'mated pair', $P = 0.419$; 'unmated pair', $P = 0.715$), the sex and the mating status of the accompanying birds modified the response of tested males to the playback of the calls of their mate (Kruskall-Wallis ANOVA, $P < 0.04$) (Fig. 3). When the male was in the presence of an established male-female pair, its vocal response was significantly stronger to its partner's calls than to the familiar female's calls (Wilcoxon matched-pairs test, $P < 0.05$). Alternatively, when the male was in the presence of unmated conspecifics, that is, two single males or one single female and one single male, its response did not differ significantly between the two stimuli sets (Wilcoxon matched-pairs test, $P = 0.593$ and 0.138 , respectively), showing no preference to the calls of its own mate. Finally, the audience did not seem to influence the male's response to the calls of a familiar female (Kruskall-Wallis ANOVA, $P = 0.348$). These experiments clearly show that male zebra finches have the ability to recognize their mate's voice, but that partner acoustic recognition is expressed differently according to the social relationships between accompanying birds. This effect of social context on acoustic communication in songbirds is in accordance with neurophysiological data: birds recorded in isolation or in visual and auditory contact with a female show different levels of electrophysiological activity²¹ and ZENK gene expression²² within auditory brain nuclei.

The present results support the idea that songbirds are capable of behaviours that are not just "incidentally communicative"^{12,13}. In spite of the fact that recognizing one's partner is often crucial to reproductive success, the vocal response to one's mate does not seem to be reflexive but can be modulated by the presence of companions. Moreover, this precise regulation of signalling behaviour with regard to the mating status of conspecifics provides evidence that songbirds experiencing gregarious life can show awareness of social relationships comparable to those demonstrated by social mammals such as primates⁷. This ability might be linked to the cognitive demands of social living^{3,4}, which are potentially an important force driving the evolution of intelligence. Indeed, one possible ultimate explanation of why male songbirds may respond to their partner's voice in the presence of a paired audience might be that the presence of a mated male expressing female-guarding behaviour could stimulate similar mate-guarding signalling by the tested male. Finally, our work emphasizes that playback results can be influenced by the presence of a particular audience, stressing the importance of considering the social environment during experiments. □

Methods

Adult zebra finches (*Taeniopygia guttata*) served as the subjects for this study and were naive to all testing procedures. Birds were bred in an aviary (12 h light/12 h dark photoperiod with adapted wavelengths); food and water were provided *ad libitum* and temperature was maintained between 23 and 25 °C. All birds were paired for several months and raised at least one brood. All experiments occurred between 9 a.m. and noon. During testing periods, temperature, food and water conditions were the same as in the aviary. The experimental protocols were approved by the Jean Monnet University's animal care committee.

Seven female zebra finches were recorded for analysis of female distance calls. Each female was isolated from its mate and seven to ten distance calls were recorded with a Sennheiser MD42 microphone, placed 0.3 m above the cage, connected to a Marantz PDM690/W1B recorder with 22,050-Hz sampling rate. We analysed a total of 65 female distance calls using Syntana software²³ and Praat version 4.0.19. We defined seventeen spectral, temporal and amplitude acoustic cues to describe the calls' acoustic structure. These measured parameters allowed statistical analysis of the cues potentially supporting individual identity coding and thus recognition by males. We performed a non-parametric analysis of variance (Kruskall-Wallis ANOVA, $P = 0.05$) and a principal components

analysis, followed by a discriminant analysis. To describe the intra-individual and inter-individual variations of each parameter we used the coefficient of variation (CV)²⁴. For each parameter we calculated CV_i (within individual CV) and CV_b (between individual CV) according to the formula for weak samples: $CV = \{100(SD/X_{mean})[1 + 1/(4n)]\}$ where SD is standard deviation, X_{mean} is the mean of the sample and n is the population sample²⁵. To assess the potential of individual coding (PIC) for each parameter, we calculated the ratio CV_b/CV_i , $mean$ (CV_i , $mean$ being the mean value of the CV_i of all individuals). For a given parameter, a PIC value greater than one suggests that this parameter may be used for individual recognition because its intra-individual variability is smaller than its inter-individual variability.

A synthetic copy of three distance calls of each female was created with Avisoft-SAS Lab Pro software (version 4.16, 2002) in order to have experimental signals with neither any background noise nor sound degradation. These synthetic calls were used for playback tests. For playback tests, each tested male bird ($n = 15$) was separated from its mate and placed in an experimental cage (240 × 50 × 50 cm, equipped with roosts) one night before the start of stimulus presentation. The experimental cage was in a soundproof chamber with a 12 h light/12 h dark photoperiod. Another cage was placed near the experimental cage in the chamber: this companion cage contained the two birds defining the social context. This audience was composed of different individuals during each trial. The playback equipment was constituted by two high-fidelity speakers (JBL TLX 12) connected to a DAT recorder (Sony DTC-ZE 700) placed at either end of the cage. During each test only one randomly chosen speaker emitted the playback stimuli (sound level, 70 dB at 1 m). The tested male was presented with two sets of stimuli: a series of distance calls of its mate and a series of distance calls of a familiar female (series were broadcast at random; series duration, 5.0 s, one call per second; interval between series, 30 s). As all the birds (audience birds, tested males and their mates, familiar females) had been bred in the same aviary, all acoustic stimuli can be considered as perceptually equivalent for the audience birds. The vocal and locomotor activity of the bird was recorded with a video recorder (Sony DCR-TRV24E) and the number of distance calls emitted during the broadcasting of the stimuli was counted. The audience behaviour was similarly assessed. Before the playback broadcasting, the male 'spontaneous' activity (call rate per 5.0 s) was also measured. The effects of the mating status of the accompanying birds on the 'spontaneous' activity of males and on the response of males to the playback of the calls of their mate or of a familiar female were analysed with Kruskal-Wallis ANOVAs ($P = 0.05$). Complementary tests (Wilcoxon matched-pairs test²⁴, $P = 0.05$) were performed to compare the responses to both stimuli in each social context. We also tested whether the audience behaviour was independent of which female's calls were being presented (Wilcoxon matched-pairs test, $P = 0.05$).

Received 31 March; accepted 11 May 2004; doi:10.1038/nature02645.

1. Barton, R. A. & Dunbar, R. I. M. in *Machiavellian Intelligence II: Extensions and Evaluations* (eds Whiten, A. & Byrne, R. W.) 240–263 (Cambridge Univ. Press, Cambridge, 1997).
2. Byrne, R. W. & Whiten, A. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans* (Clarendon, Oxford, 1988).
3. Dunbar, R. I. M. The social brain hypothesis. *Evol. Anthropol.* **6**, 178–190 (1998).
4. Adolphs, R. Cognitive neuroscience of human social behaviour. *Nature Rev. Neurosci.* **4**, 165–178 (2003).
5. Perry, S., Barrett, C. & Manson, J. H. White-faced capuchin monkeys show triadic awareness in their choice of allies. *Anim. Behav.* **67**, 165–170 (2004).
6. Jolly, A. Lemur social behaviour and primate intelligence. *Science* **153**, 501–506 (1966).
7. Tomasello, M. & Call, J. *Primate Cognition* (Oxford Univ. Press, Oxford, 1997).
8. Baltz, A. P. & Clark, A. B. Extra-pair courtship behaviour of male budgerigars and the effect of an audience. *Anim. Behav.* **53**, 1017–1024 (1997).
9. McGregor, P. K., Otter, K. & Peake, T. M. in *Animal Signals: Signalling and Signal Design in Animal Communication* (eds Espmark, Y., Amundsen, T. & Rosenqvist, G.) 329–340 (Tapir Academic, Trondheim, Norway, 2000).
10. Cynn, J. & Gell, C. Social mediation of vocal amplitude in a songbird, *Taeniopygia guttata*. *Anim. Behav.* **67**, 451–455 (2004).
11. Zann, R. *The Zebra Finch: A Synthesis of Field and Laboratory Studies* 196–213 (Oxford Univ. Press, Oxford, 1996).
12. Evans, C. S. & Evans, L. Chicken food calls are functionally referential. *Anim. Behav.* **58**, 307–319 (1999).
13. Evans, C. S. & Marler, P. Food calling and audience effects in male chickens, *Gallus gallus*: their relationships to food availability, courtship and social facilitation. *Anim. Behav.* **47**, 1159–1170 (1994).
14. Marler, P., Duffy, A. & Pickert, R. Vocal communication in the domestic chicken: II. Is a sender sensitive to the presence and nature of a receiver? *Anim. Behav.* **34**, 194–198 (1986).
15. Striedter, G. F., Freibott, L., Hile, A. G. & Burley, N. T. For whom the male calls: an effect of audience on contact call rate and repertoire in budgerigars, *Melospittacus undulatus*. *Anim. Behav.* **65**, 875–882 (2003).
16. Johnstone, A. R. Eavesdropping and animal conflict. *Proc. Natl Acad. Sci. USA* **98**, 9177–9180 (2001).
17. Miller, D. B. The acoustic basis of mate recognition by female zebra finches (*Taeniopygia guttata*). *Anim. Behav.* **27**, 376–380 (1979).
18. Charrier, L., Mathevon, N. & Jouventin, P. Mother's voice recognition by seal pups. *Nature* **412**, 873 (2001).
19. Belin, P., Fecteau, S. & Bédard, C. Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* **8**, 129–135 (2004).
20. Insley, S. J. Long-term recognition in the northern fur seal. *Nature* **406**, 404–405 (2000).
21. Hessler, N. A. & Doupe, A. J. Social context modulates singing-related neural activity in the songbird forebrain. *Nature Neurosci.* **2**, 209–211 (1999).
22. Jarvis, E. D., Scharff, C. & Grossman, M. R. For whom the bird sings: context-dependent gene expression. *Neuron* **21**, 775–788 (1998).
23. Aubin, T. Syntana: a software for the synthesis and analysis of animal sounds. *Bioacoustics* **6**, 80–81 (1994).
24. Sokal, R. R. & Rohlf, F. J. *Biometry* 57–58 & 443 (Freeman, New York, 1995).

25. Robinson, P., Aubin, T. & Brémond, J. C. Individuality in the voice of emperor penguin *Aptenodytes forsteri*: adaptation to a noisy environment. *Ethology* **94**, 279–290 (1993).

Acknowledgements This research was supported by the Interdisciplinary Program “Cognition and Information Processing” of the Centre National de la Recherche Scientifique (CNRS). C.V. is supported by the French Ministry of National Education. We thank C. Bouchut, M. Rabearinaivo and S. Tisseur for their help during the work, L. Bloomfield for improving the English, and I. Charrier for her advice on previous drafts of the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.M. (mathevon@univ-st-etienne.fr).

Anthrax kills wild chimpanzees in a tropical rainforest

Fabian H. Leendertz^{1,2,3}, Heinz Ellerbrok², Christophe Boesch¹, Emmanuel Couacy-Hymann⁴, Kerstin Mätz-Rensing⁵, Regine Hakenbeck⁶, Carina Bergmann⁶, Pola Abaza^{1,2}, Sandra Junglen^{1,2}, Yasmin Moebius¹, Linda Vigilant¹, Pierre Formenty⁷ & Georg Pauli²

¹Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany

²Zentrum für biologische Sicherheit, Robert Koch-Institut, Nordufer 20, D-13353 Berlin, Germany

³Institute for Parasitology and International Animal Health, Free University of Berlin, Königsplatz 67, D-14163 Berlin, Germany

⁴Lanada/Lcpa, Bingerville, Ivory Coast

⁵German Primate Center, Kellnerweg 4, D-37077 Göttingen, Germany

⁶Department of Microbiology, University of Kaiserslautern, Paul-Ehrlich-Straße, D-67663 Kaiserslautern, Germany

⁷Ebola Tai Forest Project, World Health Organisation, WHO Office, Abidjan, Ivory Coast

Infectious disease has joined habitat loss and hunting as threats to the survival of the remaining wild populations of great apes. Nevertheless, relatively little is known about the causative agents^{1–3}. We investigated an unusually high number of sudden deaths observed over nine months in three communities of wild chimpanzees (*Pan troglodytes verus*) in the Taï National Park, Ivory Coast. Here we report combined pathological, cytological and molecular investigations that identified *Bacillus anthracis* as the cause of death for at least six individuals. We show that anthrax can be found in wild non-human primates living in a tropical rainforest, a habitat not previously known to harbour *B. anthracis*. Anthrax is an acute disease that infects ruminants^{4,5}, but other mammals, including humans, can be infected through contacting or inhaling high doses of spores or by consuming meat from infected animals⁶. Respiratory and gastrointestinal anthrax are characterized by rapid onset, fever, septicaemia and a high fatality rate without early antibiotic treatment^{6,7}. Our results suggest that epidemic diseases represent substantial threats to wild ape populations, and through bushmeat consumption also pose a hazard to human health.

Long-term studies of wild chimpanzees habituated to human observation have revealed that mortality rates of adult animals are high and several times in excess of those reported for human hunter-gatherer communities⁸. Unfortunately, although direct observation of ill animals has suggested infectious disease as the cause of many deaths and disappearances, practical difficulties concerning acquisition and testing of diagnostic samples have usually precluded identification of the disease-causing agent(s). However, one recent

study implicated Ebola virus as the cause of a more than 50% reduction in the number of chimpanzees and gorillas in large areas of Central Africa⁹. Ebola was also shown as the main cause of death in two epidemics in 1992 and 1994 that reduced the size of a monitored community of chimpanzees in the Taï National Park from 51 to 31 individuals^{10,11}.

The behaviour and apparent health of chimpanzees in the Taï National Park have been closely monitored since 1984, and three communities (North, Middle and South) are currently under observation¹¹. In October 2001, four apparently healthy individuals (Dorry, Gargantua, Gisèle and Goma) of the North community left the main group and were not followed by observers. Three days later, the one- to two-day-old remains of three of the chimpanzees were found within 50 metres of each other. Six weeks later, a skeleton attributed to Goma was found about 200 metres away.

A new cluster of unexpected deaths occurred in the Middle community in February 2002. The body of Noah, a juvenile male who exhibited no previous signs of illness, was found the morning of 13 February near his sleeping nest, and was estimated to have been dead for approximately five hours. The next day, Léo, the top-ranking male of the same community, showed sudden signs of weakness, vomited several times and died within two hours of the onset of symptoms. Another community member, Koulo, was last observed on that day and never seen again despite intensive searching, and hence is presumed dead. Finally, in June 2002, Olduvai, a subadult male of the South community, was found near to death less than three hours after he had been last observed in normal condition.

In total, eight sudden deaths were recorded in the three chimpanzee communities between October 2001 and June 2002, and all of the individuals were in apparent good health shortly before, suggesting an acute infectious agent as the cause. Samples from all six available dead individuals tested negative for the filo- and arenaviruses Ebola, Marburg and Lassa. Pathological and histological examination of the remains of Léo and Noah revealed haemorrhages presenting as small ecchymoses in nearly all inner organs, particularly in the intestines and lungs, and the lungs were also characterized by oedema and emphysema. Microscopic examination revealed Gram-positive, rod-shaped bacteria located intra- and extravascularly in all tissues examined—spleen, liver, lung,

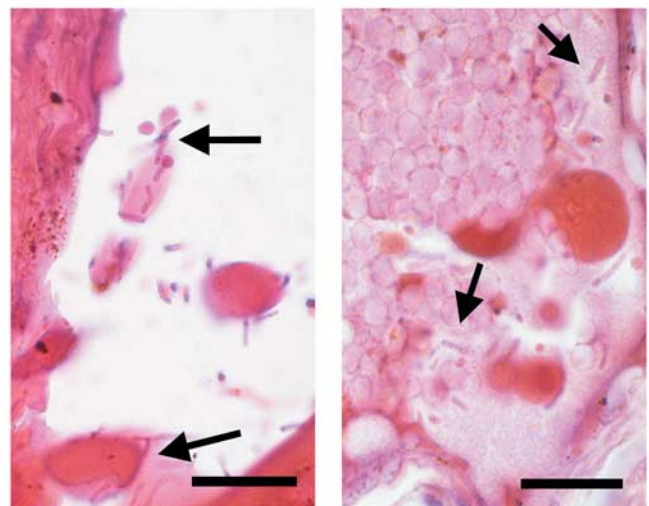


Figure 1 Histological section of lung tissues of the chimpanzee Léo. Thin sections of altered tissues were stained with haematoxylin and eosin. Under a light microscope, long rod-shaped bacteria (arrows) were visible. Left panel, section of the lung parenchyma with oedema and emphysema. Right panel, intravascular bacteria. Scale bars, 20 µm.

lymph nodes, intestines (Fig. 1), suggesting an acute bacterial infection as the cause of death. Specific amplification and sequencing of bacterial 16S RNA genes revealed identity to two closely related bacteria, *Bacillus cereus* (GenBank accession number AY138279) and *B. anthracis* (GenBank accession number AE017025). A real-time polymerase chain reaction (PCR) assay targeting *B. anthracis*-specific *pag* and *capC* virulence genes encoded by the plasmids pXO1 and pXO2, respectively, and the chromosomal *rpoB* gene¹², was performed. All six of the dead individuals sampled were positive for *B. anthracis*. Variable-number tandem repeat (VNTR) analysis detected the presence of two repeat units in the previously described *vrA* region^{13,14}. This particular allele was identified in all six animals, indicating that they had probably been infected with the same strain of *B. anthracis*. Preliminary results with pXO1-aat and pXO2-at, two VNTRs with high discriminatory power, were compatible with an African origin of the *B. anthracis* strain¹⁵.

These molecular findings, and the agreement of the pathological and cytological findings with those reported in experimental infections of non-human primates with anthrax¹⁶, suggest that all of the six individuals tested died of anthrax. Circumstantial evidence suggests that the other two individuals (Goma, Koulo), who disappeared at the same time but could not be tested, also died of anthrax. None of the DNA samples obtained from six additional chimpanzees that had died before 2001, nor samples from Kady, who died in 2001 after a long illness, were positive for *B. anthracis* (see Supplementary Information).

There are several plausible means by which the chimpanzees could have become infected with anthrax. One is through the consumption of an infected animal, such as the ungulate herbivores most often afflicted. However, detailed observation of chimpanzee hunting behaviour in the Tai National Park for more than 16 years has revealed that the chimpanzees ignore even easily available antelopes and prey exclusively upon various monkey species¹¹.

A second possibility is the ingestion of spores from contaminated water⁷. This would be supported by anthrax deaths of other species, but extensive searches revealed no die-offs of other animals. However, there are many different water sources available in Tai National Park, and limited average visibility in the forest means individual carcasses could have easily been missed. However, the significance of the source of infection is considerable. Contaminated water might have a substantially greater long-term impact than the consumption of a single infected animal. Isolated cases of anthrax have been imported to other parts of Ivory Coast by animal transports from countries where anthrax is endemic¹⁷. Indeed, owing to deforestation, in recent years cattle transports from Mali and Burkina Faso have passed close to the border of the Tai National Park, but at present there is no evidence of an introduction of anthrax by this route.

Regardless of the source of infection, the occurrence of anthrax in non-human primates is of concern not only as a threat to the survival of highly endangered species, but as a potential risk to human health. An increasing amount of evidence suggests that viral pathogens such as STLV-1, SIV and Ebola move readily between primate species^{18,19}, including humans^{3,9,20}. The hunting and consumption of bushmeat clearly has the potential to expose humans to a wide range of deadly pathogens and should be strongly discouraged. □

Methods

Sample collection and physical examination

Multiple tissue samples were collected from all chimpanzee remains and were preserved both in liquid nitrogen and by immersion in 10% buffered formaldehyde. Because of fast progressive autolysis under tropical conditions, pathological and histopathological examinations of the sudden death cases were only possible for Léo and Noah. For histological examination, the fixed paraffin-embedded tissues were processed and stained with either haematoxylin and eosin, or PAS (periodic acid Schiff) reaction and Giemsa.

Molecular analysis

DNA was extracted from various tissues following the standard protocol of the DNeasy tissue kit (Qiagen). DNA was prepared on separate days and samples were treated from one animal at a time. DNA was stored in aliquots at -20°C . The identities of the remains attributed to Gisèle and Gargantua were confirmed by comparison of microsatellite genotypes from these remains with those previously produced from these individuals²¹. DNA representing 16S RNA sequences was amplified by PCR with bacterial-specific primers. Three different real-time PCRs targeting the *B. anthracis*-specific *pag* and *capC* genes encoded by the plasmids pXO1 and pXO2, respectively, and the chromosomal *rpoB* gene were performed¹². Each amplification was attempted a minimum of two times from each sample. All samples that were positive in histological analyses were always positive for anthrax in all three PCR assays, whereas samples that were negative in histological analysis were also PCR negative. Plasmids pXO1 and pXO2 both need to be present in pathogenic *B. anthracis* but are absent in other closely related bacteria. They can therefore be used to differentiate virulent *B. anthracis* from apathogenic variants and the closely related *Bacillus cereus*, *Bacillus thuringiensis*, and *Bacillus megaterium*¹². PCR products were sequenced on both strands using the corresponding PCR primers and compared to published sequences in the public databases (GenBank, EMBL-Bank). PCR amplification of a VNTR region of the *vrA* gene was performed as described¹³. Tests for Ebola, Marburg, and Lassa virus, using highly sensitive real-time PCR assays²², were conducted at the Bernhard-Nocht-Institut, Hamburg.

Received 28 February; accepted 7 June 2004; doi:10.1038/nature02722.

1. Wolfe, N. D. *et al.* Wild primate populations in emerging infectious disease research: The missing link? *Emerg. Infect. Dis.* **4**, 149–158 (1998).
2. Ferber, D. Human diseases threaten great apes. *Science* **289**, 1277–1278 (2000).
3. Walsh, P. D. *et al.* Catastrophic ape decline in western equatorial Africa. *Nature* **422**, 611–614 (2003).
4. Smith, K. L. *et al.* *Bacillus anthracis* diversity in Kruger National Park. *J. Clin. Microbiol.* **38**, 3780–3784 (2000).
5. Turnbull, P. C. B., Doganay, M., Lindeque, P. M., Aygen, B. & McLaughlin, J. Serology and anthrax in humans, livestock and Etosha National Park wildlife. *Epidemiol. Infect.* **108**, 299–313 (1992).
6. Sirisanthana, T. & Brown, A. E. Anthrax of the gastrointestinal tract. *Emerg. Infect. Dis.* **8**, 649–651 (2002).
7. Beatty, M. E., Ashford, D. A., Griffin, P. M., Tauxe, R. V. & Sobel, J. Gastrointestinal anthrax: review of the literature. *Arch. Intern. Med.* **163**, 2527–2531 (2003).
8. Hill, K., Boesch, C., Goodall, J., Pusey, A., Williams, J. & Wrangham, R. Mortality rates among wild chimpanzees. *J. Hum. Evol.* **40**, 437–450 (2001).
9. Leroy, E. M. *et al.* Multiple Ebola virus transmission events and rapid decline of central African wildlife. *Science* **303**, 387–390 (2004).
10. Formenty, P. *et al.* Ebola virus outbreak among wild chimpanzees living in a rain forest of Côte d'Ivoire. *J. Infect. Dis.* **179** (Suppl. 1), 120–126 (1999).
11. Boesch, C. & Boesch-Achermann, H. *The Chimpanzees of the Tai Forest: Behavioural Ecology and Evolution* (Oxford Univ. Press, Oxford, 2000).
12. Ellerbrok, H. *et al.* Rapid and sensitive identification of pathogenic and apathogenic *Bacillus anthracis* by real-time PCR. *FEMS Microbiol. Lett.* **214**, 51–59 (2002).
13. Andersen, G. L., Simchok, J. M. & Wilson, K. H. Identification of a region of genetic variability among *Bacillus anthracis* strains and related species. *J. Bacteriol.* **178**, 377–384 (1996).
14. Jackson, P. J. *et al.* Characterisation of the variable-number tandem repeats in *vrA* from different *Bacillus anthracis* isolates. *Appl. Environ. Microbiol.* **63**, 1400–1405 (1997).
15. Keim, P. *et al.* Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. *J. Bacteriol.* **182**, 2928–2936 (2000).
16. Vasconcelos, D. *et al.* Pathology of inhalation anthrax in cynomolgus monkeys (*Macaca fascicularis*). *Lab. Invest.* **83**, 1201–1209 (2003).
17. Boutin, J. P., Debonne, J. M. & Rey, J. L. Apparition du charbon humain en forêt Ivoirienne. *Méd. Tropicale* **45**, 79–81 (1985).
18. Leendertz, F. H. *et al.* High variety of different STLV-1 strains in chimpanzees (*Pan troglodytes verus*) of the Tai National Park, Côte d'Ivoire. *J. Virol.* **78**, 4352–4356 (2004).
19. Bailes, E. *et al.* Hybrid origin of SIV in chimpanzees. *Science* **300**, 1713 (2003).
20. Hahn, B., Shaw, G. M., De Cock, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–614 (2000).
21. Vigilant, L., Hofreiter, M., Siedel, H. & Boesch, C. Paternity and relatedness in wild chimpanzee communities. *Proc. Natl Acad. Sci. USA* **98**, 12890–12895 (2001).
22. Drosten, C. *et al.* Rapid detection and quantification of RNA of Ebola and Marburg viruses, Lassa virus, Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus, dengue virus, and yellow fever virus by real-time reverse transcription-PCR. *J. Clin. Microbiol.* **40**, 2323–2330 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Ivorian authorities for long-term support, especially the Ministry of the Environment and Forests as well as the Ministry of Research, the directorship of the Tai National Park, and the Swiss Research Center in Abidjan. We are grateful to P. Emerich and H. Schmitz at the Bernhard-Nocht-Institut for VHF analyses. For technical support we thank S. Yumlu, T. Deschner, E. Normand, S. Pociuli, H. Emmel, and also U. Erikli for copy editing. This work was supported by the Robert Koch-Institut and the Max-Planck-Gesellschaft.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.E. (ellerbrok@rki.de).

Conceptual precursors to language

Susan J. Hespos¹ & Elizabeth S. Spelke²

¹Department of Psychology and Human Development, Vanderbilt University, Nashville, Tennessee 37203, USA

²Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA

Because human languages vary in sound and meaning, children must learn which distinctions their language uses. For speech perception, this learning is selective: initially infants are sensitive to most acoustic distinctions used in any language^{1–3}, and this sensitivity reflects basic properties of the auditory system rather than mechanisms specific to language^{4–7}; however, infants' sensitivity to non-native sound distinctions declines over the course of the first year⁸. Here we ask whether a similar process governs learning of word meanings. We investigated the sensitivity of 5-month-old infants in an English-speaking environment to a conceptual distinction that is marked in Korean but not English; that is, the distinction between 'tight' and 'loose' fit of one object to another^{9,10}. Like adult Korean speakers but unlike adult English speakers, these infants detected this distinction and divided a continuum of motion-into-contact actions into tight- and loose-fit categories. Infants' sensitivity to this distinction is linked to representations of object mechanics¹¹ that are shared by non-human animals^{12–14}. Language learning therefore seems to develop by linking linguistic forms to universal, pre-existing representations of sound and meaning.

Our research focuses on the crosscutting conceptual distinctions between actions producing loose- and tight-fitting contact relationships (compare left and right columns in Fig. 1a) and actions producing containment versus support relationships (compare first and second rows in Fig. 1a). As early as Korean and English children begin to talk about such actions, they categorize them differently from one another and similarly to Korean- and English-speaking adults^{9,15}. Moreover, English and Korean adults differ in their performance on non-linguistic categorization tasks involving heterogeneous examples of these actions, in accord with the differing semantics of their languages^{16,17}, whereas the performance of young children on such tasks has been mixed^{9,10,18,19}. These findings suggest that learning the semantics of a natural language influences one's conceptualization of the world^{20,21}, but what is the nature of this influence? It is possible that language learning creates new conceptual categories: by hearing the expression 'put on' applied to the actions of placing a book on a table or a ring on a finger, for example, speakers of English may come to perceive similarities among these events^{9,21}. Alternatively, sensitivity to conceptual distinctions that are central to the semantics of any human language may emerge before language experience and then be enhanced or diminished by subsequent experience¹⁰. To investigate these possibilities, we tested the sensitivity of infants living in a monolingual English environment to the conceptual distinction between actions that create tight- and loose-fitting contact relationships, both within and across the English-marked distinction between containment (in) and support (on).

The experiments used a looking time procedure, relying on infants' tendency to habituate to repeated events and look longer at novel ones²². Experiment 1 investigated whether infants show categorical perception of tight- and loose-fitting actions, as they do for speech sounds not present in their native language³. Five-month-old infants were presented with a continuum of events in which a person put a cylindrical object into a cylindrical container that held it loosely or tightly (Fig. 1a). Infants in two conditions were habituated to an event in which either a narrow cylinder (loose condition) or a wide cylinder (tight condition) was placed into a

medium-width container, and then all the infants were tested with events involving the narrow cylinder and two new containers, one narrower (tight) and one wider (loose) than the container presented during habituation. Infants looked longer at the test actions that presented a change from tight- to loose-fitting containment or vice versa ($F_{1,30} = 18.59$, $P < 0.001$; Fig. 1b). As with Korean adults in past research, infants therefore divided this continuum of events into the categories of tight-fitting and loose-fitting relationships.

a

Experiment 1

(i) Loose-in condition



(ii) Tight-in condition



Experiment 2

(iii) Loose-on condition



(iv) Tight-on condition



Test trials

(v) Loose-fit event



(vi) Tight-fit event



b

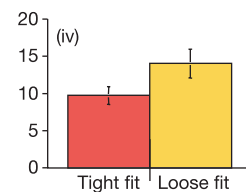
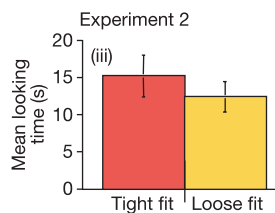
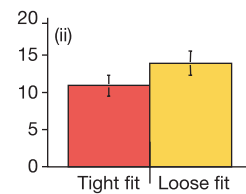
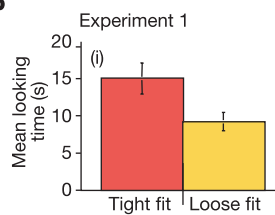


Figure 1 Infants show categorical perception of tight- and loose-fitting actions.

a, Events for experiments 1 and 2 involving habituation to loose containment (i), tight containment (ii), loose support (iii), or tight support (iv), and testing with new loose containment (v) and tight containment (vi) events. **b**, Test trial looking times in experiments 1 (top row) and 2 (bottom row). Preference for the novel relationship was significant in each experiment. Three conditions were significant when analysed separately (condition i, $F_{1,15} = 13.21$, $P < 0.01$; condition ii, $F_{1,15} = 0.51$, $P < 0.05$; condition iv, $F_{1,15} = 6.08$, $P < 0.05$); the preference was marginal in the 'loose-on' condition (condition iii, $F_{1,15} = 1.61$, $P = 0.2$). Error bars represent standard error.

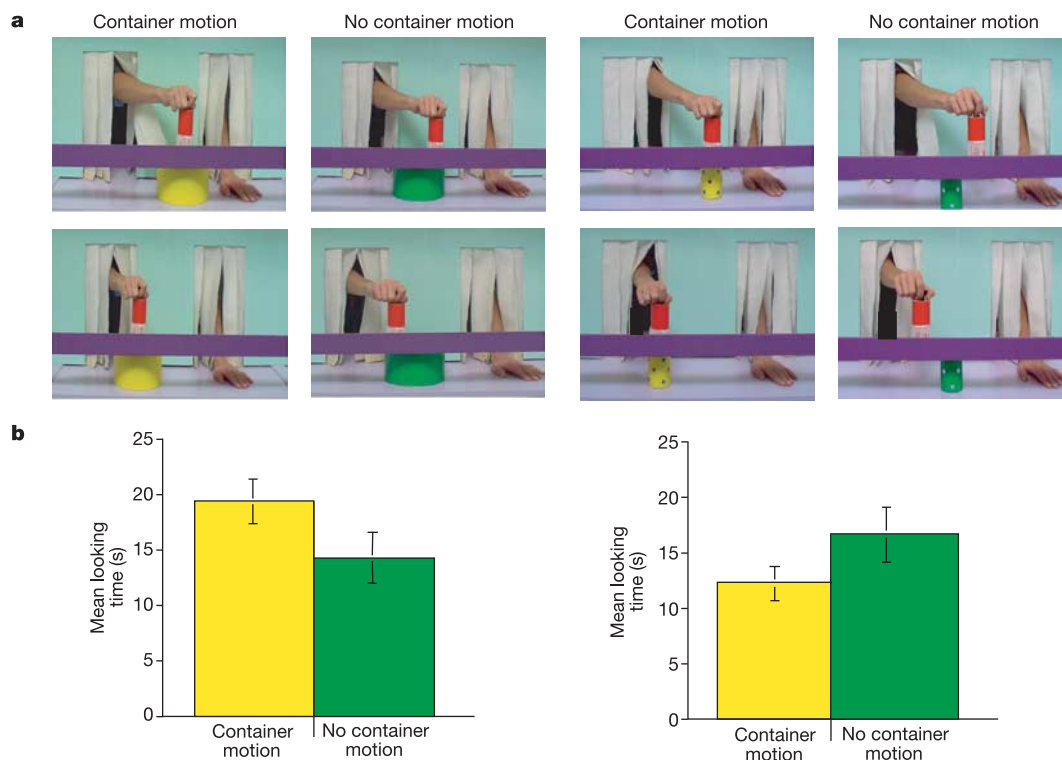


Figure 2 Infants use the tight-loose distinction in predicting object motion. **a**, Test events in experiment 3. **b**, Test trial looking times in experiment 3. Preference for the unnatural motions was significant, both overall and in each condition (each $F_{1,15} > 4.5$,

$P < 0.05$). Looking preferences in each test condition differed reliably from those of control conditions presenting the events of the test trials without exposure to the containment events. Error bars represent standard error.

Experiment 2 investigated whether infants in English-speaking families, like Korean adults, generalize the tight-loose distinction across variations in a mechanical distinction that is lexicalized in the closed-class morphology of English but not of Korean: the distinction between containment and support. Five-month-old infants in two conditions were habituated either to a loose-fitting support event, in which a solid object was placed on a pedestal, or to a tight-fitting support event, in which a hollow object was placed on a post (Fig. 1a). Then infants were tested with the same two test events as in experiment 1; that is, events that would be categorized in English as containment and in Korean as a tight or loose fit. Infants looked longer at the test events that presented a change from tight- to loose-fitting or vice versa ($F_{1,30} = 6.422$, $P < 0.02$; Fig. 1b). As with Korean adults, infants exposed only to the English language categorized actions as causing tight- versus loose-fitting contact, even when the actions crosscut the English distinction between 'put in' and 'put on'.

What is the source of these action categories? Both human infants and non-human primates represent the mechanical properties of objects by analysing the arrangements and motions of surfaces^{11–14,23}, and these relationships differ for tight- versus loose-fitting objects. When an object enters a loose-fitting container, it can move independently in the container up to the container's boundaries; however, when an object enters a tight-fitting container, any non-accidental motion of the object will induce a corresponding motion in the container. For infants, therefore, the categorical distinction between tight- and loose-fitting relationships may be a product of a more general, language-independent system for representing object mechanics. As an initial test of this possibility, experiment 3 investigated whether 5-month-old infants make contrasting inferences about the motions of objects in tight- versus loose-fitting containers. Separate groups of 5-month-old infants were habituated to an event in which an object was placed in either a tight-fitting or

loose-fitting container. Then both groups of infants were tested with events in which the contained object was moved and the container either moved with it or remained at rest (Fig. 2a). Infants presented with a loose-fitting relationship looked longer when the cylinder and its container moved together; those presented with a tight-fitting relationship showed the reverse preference ($F_{1,30} = 9.108$, $P < 0.01$; Fig. 2b). These findings provide evidence that infants use the tight-loose distinction in predicting object motion: they infer that motion of a contained object will cause a conjoint, rigid motion of the container if, and only if, the object and container fit tightly. Because non-human primates display similar capacities¹⁴, infants' action categories seem to be linked to a language-independent system for representing objects, rather than to any representation specific to the language faculty. When language evolved as a system for linking sounds and concepts, it probably built upon a repertoire of pre-existing conceptual capacities.

Young infants in an English-speaking community are predisposed to parse a continuum of actions at a boundary point that marks a semantic distinction in Korean. Because this capacity is observed well before the acquisition of a natural language in infants whose ambient language does not mark the distinction, this capacity does not depend on language experience. Instead, the capacity seems to be linked to mechanisms for representing objects and their motions that are shared by other animals and therefore evolved before the human language faculty. Finally, infants apply the tight-loose distinction to categorize actions more consistently than do English-speaking adults tested with other materials^{16,17}, and they seem to apply the distinction more clearly than adults tested with the present materials (Fig. 3). To the extent that language experience influences the prominence of this conceptual distinction, our findings suggest that the influence is selective: language experience reduces sensitivity to conceptual distinctions not marked by the

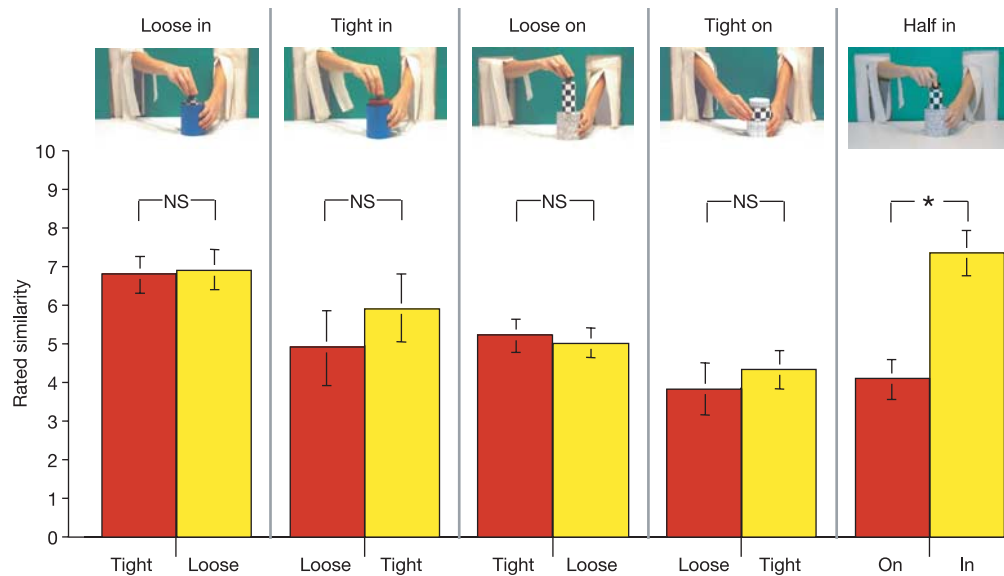


Figure 3 English-speaking adults' sensitivity to the tight-fit-loose-fit and support-containment distinctions. In judging the similarity (where 10 = high similarity) of each of two test events to a standard event (top), adults showed little sensitivity to the tight-loose distinction ($t_{1,15} < 1$ (for all t -values), not significant (NS)) and high sensitivity to the support-containment distinction ($t_{1,15} = 5.74$, $P < 0.001$). After the experiment, subjects were asked whether the actions could be grouped into two categories. All

subjects categorized the support-containment actions appropriately, and many categorized the tight-loose actions appropriately, although less consistently. The tight-loose distinction therefore seems to be accessible, on reflection, to many adults whose language does not mark it. Asterisk, $P < 0.001$. Error bars represent standard error.

native language, but it does not produce the relevant concepts. In all the above respects, the early development of semantic categories parallels the development of phonological categories and suggests that natural language semantics, like natural language phonology, evolved so as to capitalize on pre-existing representational capacities²⁴. Nevertheless, intuition suggests a difference between mature auditory and conceptual capacities. In studies of speech perception, adults' recognition of non-native phonological categories may improve with training but rarely attains native facility²⁵. In contrast, mature English speakers have little difficulty distinguishing tight-fit from loose-fit categories once these are pointed out, and many English speakers discover the categories on their own. The effects of language experience therefore may be more dramatic at the interface of audition and phonology than at the interface of conceptual structure and semantics. □

Methods

In each experiment with infants, 32 participants aged 4.5–5.5 months (16 per condition) viewed habituation and test events. Looking time was recorded from a live video by observers blind to the condition (inter-observer agreement = 92%). Each habituation or test trial continued until the infant looked away for 2 s continuously; the habituation series ended when mean looking times over triplets of trials declined by half (maximum, nine trials). Habituation trials presented a single tight- or loose-fitting event repeatedly (see Supplementary Information). Test trials presented two alternating events with initial order counterbalanced (six trials total). Test trial looking times were analysed via 2 by 3 by 2 (condition by test trial pair by test event) analyses of variance. In experiments 1 and 2, infants viewed a hand-held cylinder entering a middle-sized container that fit it either loosely or tightly (experiment 1), or placed on a support that was either loose- or tight-fitting (experiment 2). After habituation to one repeated event, all infants viewed the same tight- and loose-fitting test events with narrow and wide containers. In experiment 3, separate groups of infants viewed tight or loose containment events behind a horizontal screen that occluded the top of the container. After habituation, they viewed alternating events in which the contained object was moved and the container either moved with it or remained at rest. The unnatural tight-fit/no container motion event was produced by detaching the top half of the contained object behind the occluder. The unnatural loose-fit/container motion event was produced by securing the contained object in a concealed narrow collar within the container.

In the experiment with adults, 80 English-speaking volunteers with no knowledge of Korean, aged 18–33 yr (16 per condition), viewed one habituation event repeated six times and two test events as in the experiments with infants, and they rated the similarity of each test event to the habituation event on a ten-point scale. Four conditions presented the habituation and test events of experiments 1 and 2; a fifth condition presented a partial

containment event for habituation, in which the narrow cylinder was inserted halfway into the narrow container, and a full containment and a support event for the test, in which the same cylinder was inserted fully into or placed on top of the containing object. If adults categorize events as do infants, they should give higher similarity ratings to the same-category test event. Similarity ratings for the same-relation versus different-relation test event were compared by t -tests.

Received 4 March; accepted 4 May 2004; doi:10.1038/nature02634.

1. Jusczyk, P. W. *The Discovery of Spoken Language* (MIT, Cambridge Massachusetts, 1997).
2. Kuhl, P. K. Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *J. Phonetics* **21**, 125–139 (1993).
3. Eimas, P. D., Siqueland, E. R., Jusczyk, P. & Vigorito, J. Speech perception in infants. *Science* **171**, 303–306 (1971).
4. Hauser, M. D. *The Evolution of Communication* (MIT, Cambridge Massachusetts, 1996).
5. Wyttenbach, R. A., May, M. L. & Hoy, R. R. Categorical perception of sound frequency by crickets. *Science* **273**, 1542–1544 (1996).
6. Kuhl, P. K. & Miller, J. D. Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science* **190**, 69–72 (1975).
7. Kuhl, P. K. in *Categorical Perception: The Groundwork of Cognition* (ed. Harnad, S.) 355–386 (Cambridge Univ. Press, New York, 1987).
8. Werker, J. F. & Tees, R. C. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* **7**, 49–63 (1984).
9. Choi, S. & Bowerman, M. Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition* **41**, 83–121 (1991).
10. McDonough, L., Choi, S. & Mandler, J. M. Understanding spatial relations: Flexible infants lexical adults. *Cogn. Psychol.* **46**, 229–259 (2003).
11. Spelke, E. S. Principles of object perception. *Cogn. Sci.* **14**, 29–56 (1990).
12. Hauser, M. D. A nonhuman primate's expectations about object motion and destination: The importance of self-propelled movement and animacy. *Dev. Sci.* **1**, 31–37 (1998).
13. Santos, L. R. & Hauser, M. D. A non-human primate's understanding of solidity: Dissociations between seeing and acting. *Dev. Sci.* **5**, F1–F7 (2002).
14. Santos, L. R. 'Core knowledges': a dissociation between spatiotemporal knowledge and contact-mechanics in a non-human primate? *Dev. Sci.* **7**, 167–174 (2004).
15. Bowerman, M. in *Language and Space. Language, Speech, and Communication* (eds Bloom, P. & Peterson, M. A.) 385–436 (MIT, Cambridge Massachusetts, 1996).
16. Choi, S., McDonough, L., Bowerman, M. & Mandler, J. M. Early sensitivity to language-specific spatial categories in English and Korean. *Cogn. Dev.* **14**, 241–268 (1999).
17. Bowerman, M. & Choi, S. in *Language Acquisition and Conceptual Development* (eds Bowerman, M. & Levinson, S.) (Cambridge Univ. Press, Cambridge, UK, 2001).
18. Casasola, M., Cohen, L. B. & Chiarello, E. Six-month-old infants' categorization of containment spatial relations. *Child Dev.* **74**, 679–693 (2003).
19. Casasola, M. & Cohen, L. B. Infant categorization of containment, support and tight-fit spatial relationships. *Dev. Sci.* **5**, 247–264 (2002).
20. Bowerman, M. & Levinson, S. (eds) *Language Acquisition and Conceptual Development* (Cambridge Univ. Press, Cambridge, 2001).
21. Whorf, B. L. *Language, Thought, and Reality* (MIT, Cambridge, 1956).
22. Bornstein, M. H. in *Measurement of Audition and Vision in the First Year of Postnatal Life: A*

- Methodological Overview* (eds Gottlieb, G. & Krasnegor, N. A.) 253–300 (Westport, Connecticut, Ablex, 1985).
23. Baillargeon, R. in *Advances in Infancy Research* (ed. Lipsitt, C. R.-C. L. P.) 305–371 (Ablex, Norwood, 1995).
24. Hauser, M., Chomsky, N. & Fitch, W. T. The faculty of language: What is it, who has it, and how did it evolve? *Science* **298**, 1569–1579 (2002).
25. Werker, J. F. in *An Invitation to Cognitive Science* (eds Gleitman, L. & Liberman, M.) 87–106 (MIT, Cambridge, 1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. Blass, K. Condry, J. Goodman and L. Markson for comments and suggestions. This work was supported by grants from the NIH and NIH NRSA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.H. (s.hespos@vanderbilt.edu).

Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3

Inah Lee*, D. Yoganarasimha, Geeta Rao & James J. Knierim

Department of Neurobiology and Anatomy, W.M. Keck Center for the Neurobiology of Learning and Memory, University of Texas Medical School at Houston, PO Box 20708, Houston, Texas 77225, USA

* Present address: Center for Memory and Brain, Boston University, 2 Cummington Street, Boston, Massachusetts 02215, USA

The hippocampus, a critical brain structure for navigation, context-dependent learning and episodic memory^{1–3}, is composed of anatomically heterogeneous subregions. These regions differ in their anatomical inputs as well as in their internal circuitry⁴. A major feature of the CA3 region is its recurrent collateral circuitry, by which the CA3 pyramidal cells make excitatory synaptic contacts on each other^{4,5}. In contrast, pyramidal cells in the CA1 region are not extensively interconnected⁴. Although these differences have inspired numerous theoretical models of differential processing capacities of these two regions^{6–13}, there have been few reports of robust differences in the firing properties of CA1 and CA3 neurons in behaving animals. The most extensively studied of these properties is the spatially selective firing of hippocampal ‘place cells’¹⁴. Here we report that in a dynamically changing environment, in which familiar landmarks on the behavioural track and along the wall are rotated relative to each other^{15,16}, the population representation of the environment is more coherent between the original and cue-altered environments in CA3 than in CA1. These results demonstrate a functional heterogeneity between the place cells of CA3 and CA1 at the level of neural population representations.

Five rats implanted with multiple recording probes in CA3 and CA1 were trained to circle clockwise (CW) on a circular track in a controlled, stable environment (‘standard session’; STD in Fig. 1). Twelve tetrodes were lowered to CA3 and 6 tetrodes were lowered to CA1 (Fig. 2b). On each day of recording, three standard sessions were interleaved with two mismatch (MIS in Fig. 1) sessions, in which the local cues on the circular track were rotated anticlockwise (ACW) and the set of distal cues was rotated clockwise (CW) by an equal amount (Fig. 1). Total mismatch angles between the local and distal cue sets varied between 45°, 90°, 135° or 180°, and each rat received 4 sets of each rotation mismatch over 8 days.

Approximately 36% of the complex spike cells in CA1 (range 26–62%) and 26% of the complex spike cells in CA3 (range 22–33%) had place fields in the first standard session of the day; other cells were isolated in pre-session sleep periods but were silent or fired sparsely with no spatial selectivity. To first describe the heterogeneity of single-unit responses to the cue rotations, we categorized the individual place field responses into 5 groups (Fig. 2a). In some cases, the place cells rotated their preferred firing locations either CW or ACW, following in the direction of the distal or local cue sets, respectively. In other cases, place fields in standard sessions disappeared in mismatch sessions (‘Disappear’) or appeared only in the mismatch sessions (‘Appear’). Some place fields could not be decisively categorized into the 4 response types described above, and were categorized as ‘Ambiguous’ (for example, when a cell had a single place field in one session and multiple place fields in the other session).

There were notable differences between CA1 and CA3 in the proportions of response types (Fig. 2c; $\chi^2 = 130.8$, $P < 0.0001$). The majority of CA3 place fields (~60%) rotated on the track (ACW, $n = 221/429$; CW, $n = 33/429$), whereas only ~27% of CA1 cells responded similarly (ACW, $n = 50/349$; CW, $n = 45/349$). In contrast, most CA1 cells altered their place fields (~73%), either showing ambiguous responses ($n = 128/349$) to the changed environments or having a robust place field in only one of the two sessions (Disappear, 98/349; Appear, 28/349). Only ~40% of CA3 cells altered their place fields in these ways. These general patterns between CA1 and CA3 were observed across all rats (Supplementary Fig. 1).

To validate objectively the categorical analysis described above, a population correlation analysis was performed. The firing rate of each cell was calculated for each 1° bin of the track, thus constructing a population firing rate vector for each of the 360 bins. The firing rate vector for each bin of the standard session was then correlated with the firing rate vector for each bin of the mismatch session to produce a STD versus MIS correlation matrix (Fig. 3; Supplementary Fig. 2). For the 45° mismatch sessions, the STD versus MIS matrices for both CA1 and CA3 showed a large correlation on the diagonal, indicating that both regions maintained strong coherence in their representations between the standard and mismatch sessions with the smallest mismatch angle tested (compare with the correlation matrices between the standard 1 and standard 2 sessions). When the mismatch angles were greater than 45°, the CA1 representations lost their coherence between the standard and mismatch sessions, as the diagonal band of high correlation disappeared. In contrast, CA3 maintained a structured, diagonal band of highly correlated activity in all mismatch types. As the mismatch angle increased, the band of high correlation shifted downward, consistent with the observation that the majority of CA3 place fields rotated ACW with the local cues.

We used circular statistics¹⁷ to analyse the subgroup of cells that maintained place fields in both the standard and mismatch sessions. Figure 4a shows the amount that each place field rotated between

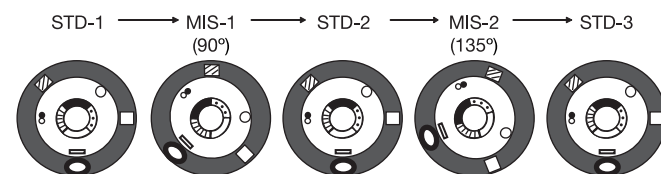


Figure 1 Experimental design. The ring track (centre) with distinctive local cues on its surface was positioned in a curtained environment (black outer circle). Distal cues were positioned along the curtained wall. Each day, the standard session (STD) was repeated three times, interleaved with cue-mismatch sessions (MIS) of different mismatch amounts (90° and 135° in this example).

the two sessions and the mean vector for each mismatch amount. The angle of the vector denotes the mean angle that the population rotated, and the length of the vector is inversely proportional to the variance of the distribution around that mean. For CA3, all mismatch amounts produced a significant vector length, indicating that the place field rotation angles were significantly clustered around the mean angle (Fig. 4a, Supplementary Fig. 3b; Rayleigh test, $P < 0.0001$). For CA1, in contrast, only the 45° and 90° mismatch sessions produced a significant vector length (Fig. 4a, Supplementary Fig. 3b; Rayleigh test, $P < 0.0001$). The CA1 cells were not randomly distributed, however (Rao's spacing test¹⁷, $P < 0.01$ for the 45°, 90° and 135° mismatch sessions; $P < 0.05$ for the 180° mismatch session); rather, there was a bimodal distribution, as roughly equivalent numbers of place fields rotated CW and ACW with their respective cue sets (Fig. 4a).

Because the preceding analyses combined data from multiple sessions over all 5 rats, the results may be an artefact of combining these different data sets. That is, it is possible that CA1 and CA3 maintained an equivalent amount of coherence in individual, simultaneously recorded data sets, but that the CA3 ensemble

always rotated ACW whereas the CA1 ensemble sometimes rotated CW and other times rotated ACW. Combining these heterogeneous data sets would lead to an incorrect conclusion that the CA3 ensemble was internally more coherent than the CA1 ensemble. To address this possibility, we calculated the mean rotation vector for each individual data set in which at least 2 cells from both CA1 and CA3 maintained a place field in the standard and mismatch sessions (for example, see data for single data sets in Fig. 4b). A Wilcoxon signed rank test (the non-parametric analogue of a paired t -test) demonstrated that on average, the mean vector length of CA3 was greater than the mean vector length of simultaneously recorded CA1 cells ($P < 0.02$), verifying that the CA3 representation was more coherent than CA1 at the level of individual ensemble recordings (Fig. 4c).

One potential explanation for the different responses of CA3 and CA1 to the altered environments is that most CA3 place cells received inputs only from neurons that represent local cues, whereas CA1 cells received inputs from neurons that represent both local and distal cues. A number of lines of evidence argue against this interpretation: (1) although the majority of CA3 place fields rotated

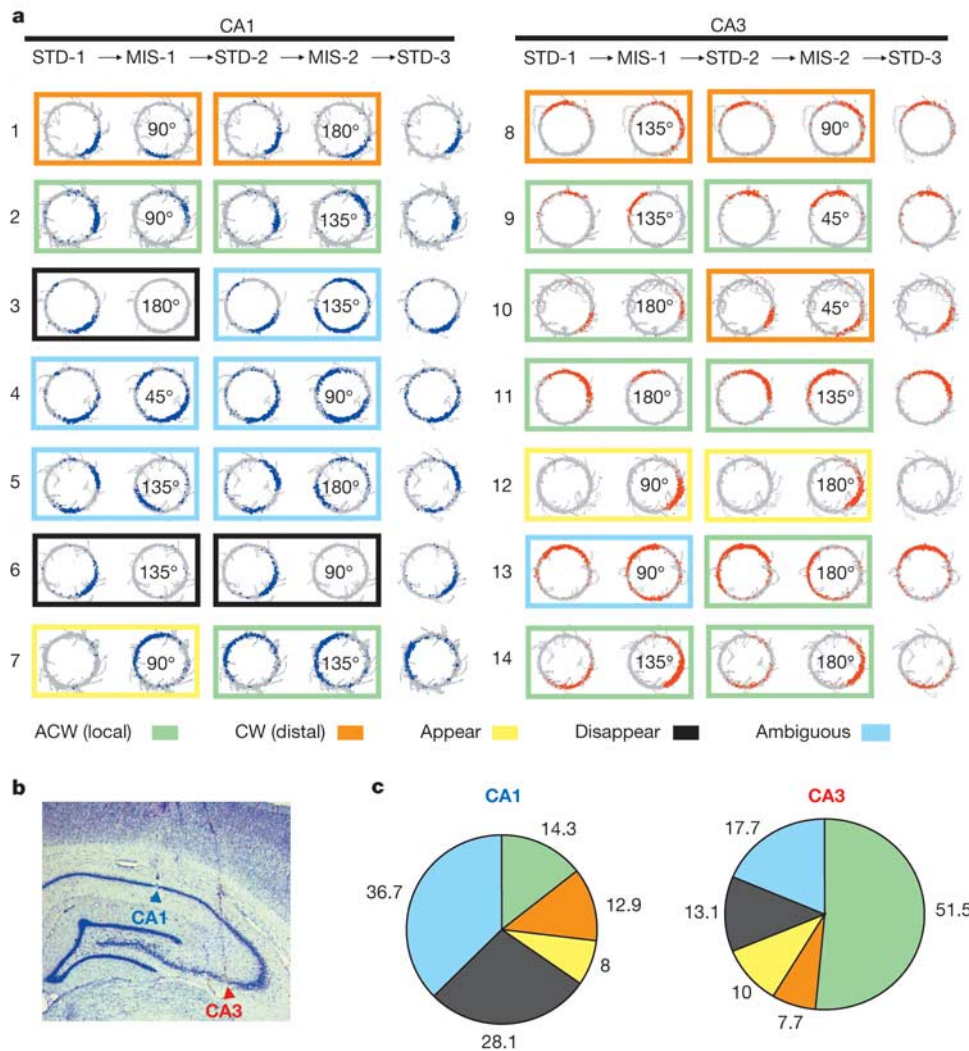


Figure 2 Individual place field responses to cue manipulations. **a**, Examples of CA1 and CA3 place fields between standard (STD) and mismatch (MIS) sessions. Each row shows five consecutive sessions in a day. Rectangles of different colours denote five response categories. Position data are shown as grey dots, and spikes as coloured dots. Angles denote mismatch amount between local and distal cue sets. These examples of place

cells were not recorded simultaneously. **b**, Representative locations of CA1 (blue arrowhead) and CA3 (red arrowhead) electrodes within a subject. **c**, Categorical classification of place field behaviour in response to mismatch sessions. Numbers show per cent.

with the local cues, almost half of the place fields rotated with the distal landmarks, remapped or behaved ambiguously. Moreover, some individual CA3 place fields were controlled by the local cues in one session and by the distal cues in another session (for example, Fig. 2a, cell 10). (2) The CA3 place fields that rotated ACW tended to rotate by a lesser amount than the local cues themselves, and this under-rotation was greater and more variable for the 180° mismatch session than for the 45° mismatch session (Fig. 4d; $P < 0.02$ between the 45° and 180° conditions, Kolmogorov-Smirnov test; see also Supplementary Fig. 4). This result demonstrates an interaction between both sets of cues in CA3. (3) A preliminary study in one of the present subjects showed that CA3 place fields were more highly correlated than CA1 place fields when distal landmarks alone were rearranged, in a situation in which the local cues were removed (Supplementary Figs 5 and 6). These data are in agreement with prior studies that showed that CA3 place fields were controlled by rotations of distal landmarks as robustly as CA1 place fields^{18,19}. Despite the evidence showing modulation of CA3 place fields by distal cues, it is still possible that CA3 is more dominantly influenced by local cues than is CA1 (especially when there is a competition between local and distal cues¹⁵). Recordings from the brain structures that send inputs to CA3 and/or CA1 (for example, the entorhinal cortex and the dentate gyrus) will be necessary to test this hypothesis rigorously. Nevertheless, the present data clearly show that neurons in CA3 respond to the altered environments in a more coherent fashion than CA1 neurons.

Since the pioneering theories of Marr⁶, the hippocampus is often modelled as an autoassociative memory network that can retrieve the originally stored representation in the presence of incomplete, corrupt, or noisy inputs ('pattern completion' or 'generalization')^{6–10,20,21}. CA3 is thought to perform such autoassociative functions via its recurrent collateral circuitry^{4–12}. Consistent with those theories, the present study provides (to our knowledge) the first direct electrophysiological evidence in normal animals

that CA3 displays far more homogeneity in its ensemble response to an altered environment than CA1. Under different circumstances, however, it might be adaptive for a network to create a new, independent representation of an altered environment⁶ (for example, if there are different behavioural contingencies associated with the familiar and altered environments). 'Pattern separation' refers to this ability of a network to make the representations of two similar input patterns more dissimilar, thus decreasing the probability of recall errors^{22–25}. Most theories consider the information processing in the dentate gyrus as an essential stage for the pattern separation process^{7,22–24}. The CA3 subfield, which receives a major input from the dentate gyrus via the mossy fibre projection, may be the region where the dynamic tension between these pattern completion and pattern separation processes is resolved^{22,23}. Under some circumstances (such as the present study, when environmental cues are altered), the CA3 network performs a pattern completion process; under other circumstances (perhaps when the spatial landmarks or other internal variables are altered to

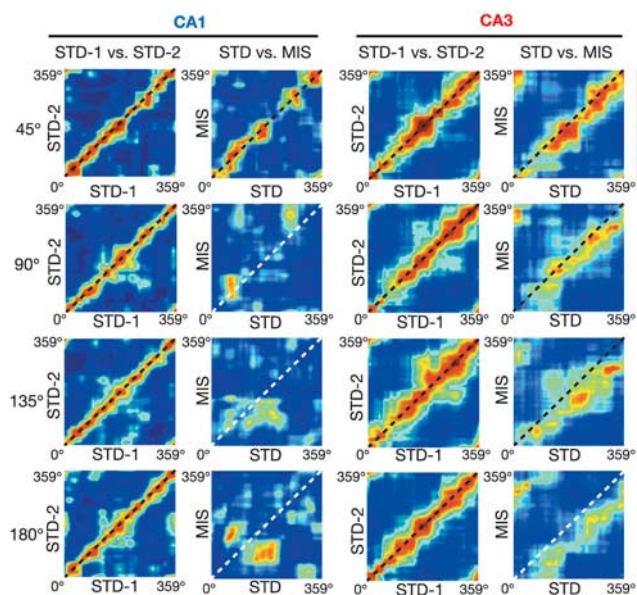


Figure 3 Correlation matrices between population firing rate vectors. Matrices shown are between population firing rate vectors of the first and the second standard sessions (STD-1 versus STD-2 correlation matrix) or between those of standard and mismatch sessions (STD versus MIS correlation matrix) for each mismatch condition. The abscissa and the ordinate represent the locations (°) on the linearized track in the standard and/or mismatch sessions. The CA3 representation remained coherent at all mismatch angles, whereas the CA1 representation abruptly lost coherence when mismatch angles exceeded 45° (see Supplementary Fig. 3a).

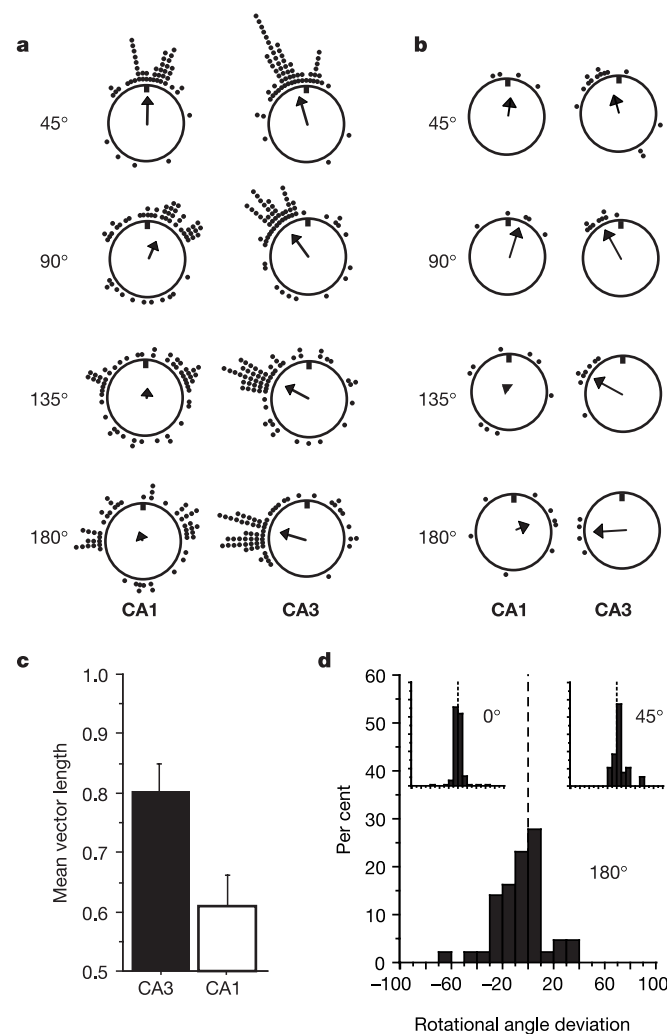


Figure 4 Ensemble coherence. **a**, Each dot represents the calculated amount of rotation of a single place field between the standard and mismatch sessions. Zero degree is denoted by the tick mark in each circle. **b**, Representative examples of ensembles of CA1 and CA3 place fields recorded simultaneously. **c**, Average vector lengths for simultaneously recorded CA1 and CA3 place field ensembles (mean $\pm$ s.e.m.). **d**, Deviation of the calculated rotation angle of the ACW-rotating CA3 place fields from the physical rotation angle of the local cue set in the 180° mismatch condition, 45° mismatch condition (right inset), and standard session (left inset). Negative numbers denote under-rotation relative to the rotation of the cue set and positive numbers denote over-rotation.

a greater extent, producing a strong pattern separation via the dentate gyrus), the CA3 ensemble transitions to a completely different location (attractor) in its representation state space^{13,25,26}.

An important finding of the present results is that the CA1 representation was not always under the strong control of its CA3 afferents^{14,25}. The role of CA1 may be to compare the rat's current experience, represented in the entorhinal inputs, with the representations of past experience stored in the DG-CA3 networks⁸. Under some circumstances, the CA1 output may be dominated by its CA3 afferents, and the CA1 and CA3 representations will behave coherently with each other (as in the 45° mismatch session; see Fig. 3). Under other circumstances, however, such as in novel environments or sufficiently altered environments (such as the larger mismatch sessions), CA1 may be dominated by its EC afferents²⁷ and its representation may be less strongly influenced by the CA3 output^{25,28}. Recordings or manipulations specifically targeting these networks in various conditions will promote further understanding of how the interactions among these networks optimize memory storage and retrieval^{10,11,22,27,29}.

Note added in proof: While our paper was in production, we became aware of another study³¹ (using immediate-early genes to measure activity) that reached conceptually similar conclusions about differences between CA1 and CA3. □

Methods

Detailed experimental procedures were described previously¹⁶. All animal procedures were in accordance with NIH guidelines on the use of experimental animals and were approved by the University of Texas Health Science Center at Houston Institutional Animal Care and Use Committee. Twelve tetrodes were lowered to CA3, and 6 tetrodes were lowered to CA1. We specifically targeted the middle or distal CA3 region because this region provides a higher density of recurrent collaterals than the proximal CA3 region⁵. We also targeted the middle and proximal CA1 region, as this part of CA1 receives topographically organized input from the middle and distal CA3, respectively^{4,5}. Only those tetrodes that allowed unambiguous assignments to either CA1 or CA3 were used for further analysis, and data from the tetrodes in CA2 or the hilar area were discarded. After a week of recovery from electrode implant surgery, the rats were trained (7–13 days) to circle clockwise (CW) on a circular track (56 cm inner diameter, 76 cm outer diameter) to collect chocolate reward placed on random locations of the track. During the course of training before recording place fields, the animals were trained to run only clockwise. Whenever the rat turned around and moved anticlockwise (ACW), its progress was stopped by blocking its path with a paper folder. The rats quickly learned to move unidirectionally in a clockwise direction. The experimenter was present in the testing room at the time of recording and, when the rat occasionally turned around to run anticlockwise, the experimenter corrected the behaviour. The track was composed of 4 different textured surfaces ('local cues'), each covering one-quarter of the ring: a grey rubber mat with a pebbled surface, brown medium-grit sandpaper, beige carpet pad material, and grey duct tape with white tape stripes. The track was placed in a circular, curtained environment (2.7 m diameter) in which six objects ('distal cues') were present either on the floor or on the curtain ('hanging cues')—a brown cardboard circle, a black and white striped card, and a white card; 'standing cues'—a white box, an intravenous stand with a laboratory coat and a blue cloth, and a roll of brown wrapping paper).

Multiple single units were recorded with the Cheetah data acquisition system (Neuralynx), and the animal's position was monitored at 30 Hz via an array of light-emitting diodes mounted on the headstage. Once the testing started, each animal was given baseline sleep periods (~30 min), during which multiple cells were recorded, before the first behavioural testing session and after the last testing session. These sleep data were used to determine the stability of recordings made during behavioural sessions and unstable cells were not further analysed.

After perfusion, electrode tracks were histologically identified and the tetrodes were carefully assigned to subfields of the hippocampus by considering both histological results and electrophysiological profiles of the tetrodes recorded during the experiments. Specifically, digital photomicrographs were taken for all serial sections of the hippocampus (40 µm). Tetrode tracks were traced on the digital images using graphic software (Adobe) and those retouched images were reconstructed three-dimensionally (Voxar). Microscopic examinations were used in parallel in the course of reconstruction. Rotated three-dimensional views of the reconstructed three-dimensional image were compared to the configuration of the tetrodes in the original tetrode-bundle for accurate identification of the tetrodes.

Units were isolated off-line using custom software for cluster cutting. The track was divided into 360 bins (1° per bin) and a firing rate for each bin was calculated by dividing the number of spikes fired while the rat occupied that bin by the amount of time spent in the bin. The place fields were smoothed by a gaussian filter with width parameter $\sigma = 4.3^\circ$. The amount that each place field rotated between the standard and mismatch sessions was calculated by computing the Pearson product-moment correlation between its firing rate array for the standard session and for the mismatch session; the latter session was shifted successively by 1° until the maximum correlation between the two arrays was found. This correlational analysis to find rotation angles was performed only when well-established

place fields (statistically significant, spatial information score³⁰ ≥ 0.5 , and number of spikes ≥ 50) were exhibited in both standard and mismatch sessions.

Received 8 January; accepted 10 June 2004; doi:10.1038/nature02739.

Published online 30 June 2004.

- O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon, Oxford, 1978).
- Eichenbaum, H. A cortical-hippocampal system for declarative memory. *Nature Rev. Neurosci.* **1**, 41–50 (2000).
- Vargha-Khadem, F. *et al.* Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**, 376–380 (1997).
- Amaral, D. G. & Witter, M. P. in *The Rat Nervous System* (ed. Paxinos, G.), 2nd edn 443–493 (Academic, San Diego, 1995).
- Ishizuka, N., Weber, J. & Amaral, D. G. Organization of intrahippocampal projections originating from CA3 pyramidal cells in the rat. *J. Comp. Neurol.* **295**, 580–623 (1990).
- Marr, D. Simple memory: a theory for archicortex. *Phil. Trans. R. Soc. Lond.* **262**, 23–81 (1971).
- Rolls, E. T. & Treves, A. *Neural Networks and Brain Function* (Oxford Univ. Press, Oxford, 1998).
- Hasselmo, M. E., Schnell, E. & Barkai, E. Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *J. Neurosci.* **15**, 5249–5262 (1995).
- McNaughton, B. L. & Morris, R. G. M. Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci.* **10**, 408–415 (1987).
- Kesner, R. P. & Rolls, E. T. Role of long-term synaptic modification in short-term memory. *Hippocampus* **11**, 240–250 (2001).
- Nakazawa, K. *et al.* Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science* **297**, 211–218 (2002).
- Treves, A. & Rolls, E. T. Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**, 374–391 (1994).
- Samsonovich, A. & McNaughton, B. L. Path integration and cognitive mapping in a continuous attractor neural network model. *J. Neurosci.* **17**, 5900–5920 (1997).
- McNaughton, B. L., Barnes, C. A. & O'Keefe, J. The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Exp. Brain Res.* **52**, 41–49 (1983).
- Shapiro, M. L., Tanila, H. & Eichenbaum, H. Cues that hippocampal place cells encode: dynamic and hierarchical representation of local and distal stimuli. *Hippocampus* **7**, 624–642 (1997).
- Knierim, J. J. Dynamic interactions between local surface cues, distal landmarks, and intrinsic circuitry in hippocampal place cells. *J. Neurosci.* **22**, 6254–6264 (2002).
- Zar, J. H. *Biostatistical Analysis*, 4th edn (Prentice Hall, Upper Saddle River, NJ, 1999).
- O'Keefe, J. & Speakman, A. Single unit activity in the rat hippocampus during a spatial memory task. *Exp. Brain Res.* **68**, 1–27 (1987).
- Cressant, A., Muller, R. U. & Poucet, B. Remapping of place cell firing patterns after maze rotations. *Exp. Brain Res.* **143**, 470–479 (2002).
- Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).
- Mizumori, S. J., McNaughton, B. L., Barnes, C. A. & Fox, K. B. Preserved spatial coding in hippocampal CA1 pyramidal cells during reversible suppression of CA3c output: evidence for pattern completion in hippocampus. *J. Neurosci.* **9**, 3915–3928 (1989).
- Kesner, R. P., Gilbert, P. E. & Wallenstein, G. V. Testing neural network models of memory with behavioural experiments. *Curr. Opin. Neurobiol.* **10**, 260–265 (2000).
- O'Reilly, R. C. & McClelland, J. L. Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* **4**, 661–682 (1994).
- McNaughton, B. L. & Nadel, L. in *Neuroscience and Connectionist Theory* (eds Gluck, M. A. & Rumelhart, D. E.) 1–63 (Erlbaum, Hillsdale, NJ, 1990).
- Leutgeb, S., Kjelstrup, K. G., Treves, A., Moser, M. & Moser, E. I. *Differential Representation of Context in Hippocampal Areas CA3 and CA1* (Program No. 91.5, Abstract Viewer/Itinerary Planner, Society for Neuroscience, Washington DC, 2003).
- Kali, S. & Dayan, P. The involvement of recurrent connections in area CA3 in establishing the properties of place fields: a model. *J. Neurosci.* **20**, 7463–7477 (2000).
- Brun, V. H. *et al.* Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. *Science* **296**, 2243–2246 (2002).
- Anderson, M. I. & Jeffery, K. J. Heterogeneous modulation of place cell firing by changes in context. *J. Neurosci.* **23**, 8827–8835 (2003).
- Lee, I. & Kenser, R. P. Differential contribution of NMDA receptors in hippocampal subregions to spatial working memory. *Nature Neurosci.* **5**, 162–168 (2002).
- Skaggs, W. E., McNaughton, B. L., Wilson, M. A. & Barnes, C. A. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**, 149–172 (1996).
- Vazdarjanova, A. & Guzowski, J. F. Differences in hippocampal neuronal population responses to modifications of an environmental context: evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *J. Neurosci.* (in the press).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This study was supported by the National Institute of Neurological Disorders and Stroke, the National Institute of Mental Health, and the Lucille P. Markey Charitable Trust. We thank E. Moser and A. Treves for comments and discussion. We also thank E. Hargreaves for assistance in collecting data for Supplementary Figs 5 and 6. G.R. and D.Y. assisted in data collection and behavioural training throughout the experiments.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to I.L. (ilee@bu.edu) or J.J.K. (James.J.Knierim@uth.tmc.edu).

Drosophila long-term memory formation involves regulation of cathepsin activity

Daniel Comas*, Florian Petit* & Thomas Preat

Développement, Évolution et Plasticité du Système Nerveux, CNRS, 1, Avenue de la Terrasse, 91190 Gif sur Yvette, France

* These authors contributed equally to this work.

Whereas short-term memory lasts from minutes to hours, long-term memory (LTM) can last for days or even an entire lifetime. LTM generally forms after spaced repeated training sessions and involves the regulation of gene expression^{1,2}, thereby implicating transcription factors in the initial steps of LTM establishment³. However, the direct participation of effector genes in memory formation has been rarely documented, and many of the mechanisms involved in LTM formation remain to be understood. Here we describe a *Drosophila melanogaster* mutant, *crammer* (*cer*), which shows a specific LTM defect. The *cer* gene encodes an inhibitor of a subfamily of cysteine proteinases, named cathepsins, some of which might be involved in human Alzheimer's disease⁴. The Cer peptide was found in the mushroom bodies (MBs), the *Drosophila* olfactory memory centre and in glial cells around the MBs. The overexpression of *cer* in glial cells but not in MB neurons induces a decrease in LTM, suggesting that Cer might have a role in glia and that the concentration of the Cer peptide is critical for LTM. In wild-type flies, *cer* expression transiently decreases after LTM conditioning, indicating that cysteine proteinases are activated early in LTM formation.

The MBs form a prominent bilateral structure of the insect brain. They are essential for short-term memory (STM), and many of the proteins involved in the early phases of memory establishment are preferentially expressed in the MBs or in neurons projecting to them⁵. We showed that MBs are also involved in LTM^{6,7}. Moreover, LTM mutants have been isolated recently⁸, but MB neuronal plasticity after LTM training remains to be characterized.

To acquire additional insights into LTM mechanisms, we have analysed *Drosophila* enhancer-trap Gal4 strains showing MB-specific expression. Flies were conditioned by exposure to an odour paired with electric shocks and subsequent exposure to a second odour in the absence of a shock. One of these strains, *MZ1180* (refs 9, 10), which we have named *crammer*^P (*cer*^P), displayed a decrease in 24-h LTM after spaced training, but a normal 24-h memory capacity after massed training, a conditioning

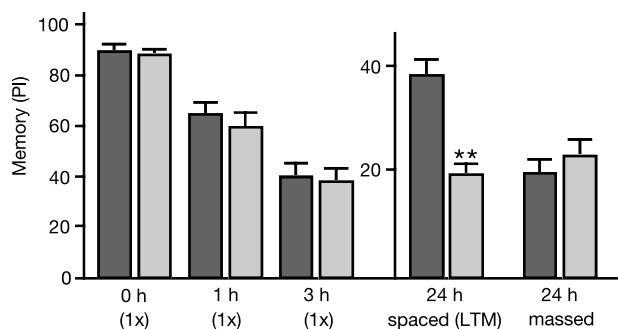


Figure 1 *cer*^P is a specific LTM mutant. Performance indices (PIs) were measured 1 and 3 h after a single conditioning cycle, and 24 h after spaced or massed training. Dark grey bars, CS; pale grey bars, *cer*^P. Results are means and s.e.m.; *n* = 8–10 groups. Asterisks indicate significant differences in a *t*-test (*P* < 0.01).

protocol that does not induce LTM² (Fig. 1). STM induced after a single training cycle was also normal, showing that *cer*^P is a specific LTM mutant. The LTM defect of *cer* is recessive (Supplementary Fig. 1). We performed excision experiments to generate revertant strains as well as additional mutant alleles. The excision line *cer*^{E29} displayed a specific LTM defect; in contrast, the *cer*^{E3} strain exhibited normal memory (Supplementary Fig. 1). Taken together, these results show that we have isolated a new LTM mutant with a P-element insertion.

Polymerase chain reaction (PCR) rescue experiments revealed that the P-element in *cer*^P flies was inserted 211 base pairs (bp) upstream of the gene encoding the succinyl-coenzyme A synthetase flavoprotein subunit (*Scs-fp*) and 53 bp downstream of the *CG10460* gene (Fig. 2a). Quantitative reverse transcriptase PCR (RT-PCR) (Fig. 2b) and northern blot (data not shown) analyses were performed to determine which gene was affected by the P-element insertion. Expression of *scs-fp* was normal in *cer*^P mutants, but expression of *CG10460* was strongly reduced, indicating that this gene corresponds to *cer* (Fig. 2b). Similar results were obtained by western blot analysis with a polyclonal antibody raised

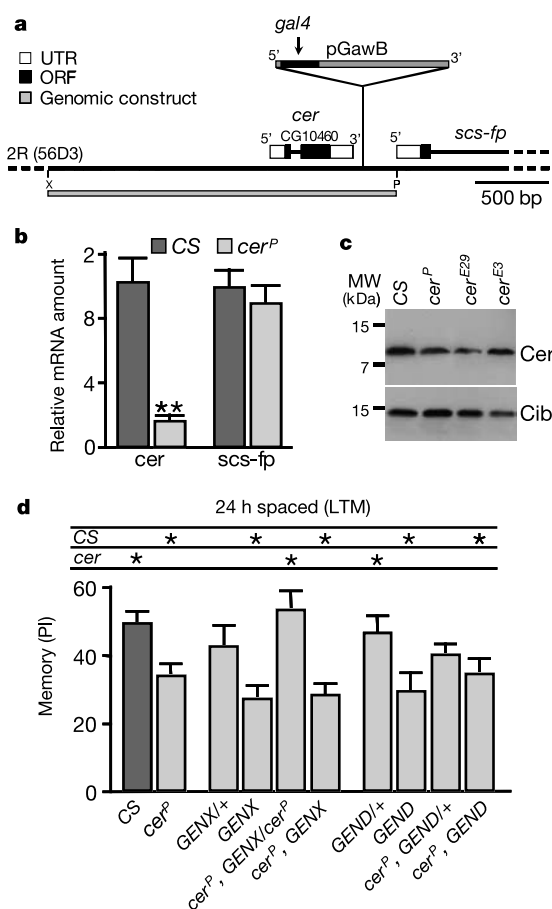


Figure 2 *cer* encodes a cathepsin inhibitor. **a**, The Gal4 P-element in *cer*^P flies is inserted between the *cer* and *scs-fp* genes. The genomic fragment used for the rescue experiment is shown with a grey bar. X, *Xba*I; P, *Pst*I; ORF, open reading frame; UTR, untranslated region. **b**, *cer* mRNA expression is affected by the P-element insertion, as measured by quantitative RT-PCR. Results are means and s.e.m.; *n* = 10–12 total RNA extractions. **c**, Similar results were obtained in a western blot assay. Densitometry analysis yielded Cer/Ciboulot (Cib) ratios of CS = 0.78 ± 0.24, *cer*^P = 0.37 ± 0.11, *cer*^{E3} = 1.29 ± 0.27 and *cer*^{E29} = 0.32 ± 0.05 (*n* = 5). **d**, The mutant phenotype can be rescued by the insertion of one genomic copy of the wild-type *cer*⁺ gene. The two upper lines represent significant differences between each score and the scores of CS or *cer*^P. PI, performance index. Results are means and s.e.m.; *n* = 6–10 groups. Asterisk indicates significant differences in a *t*-test (*P* < 0.05).

against Cer (Fig. 2c). As expected, a decreased level of Cer was found in the mutant excision *cer^{E29}*, and a normal level was found in the revertant excision *cer^{E3}* (Fig. 2c).

Sequence comparisons showed that Cer is highly similar to the amino-terminal regions of cathepsins L, a subfamily within the papain-like cysteine proteinase family (Supplementary Fig. 2a). Cysteine proteinases, which are generally synthesized as inactive proenzymes with an inhibitory N-terminal propeptide region, range in size from 300 to 350 amino acids. Interestingly, Cer is only 79 amino acids long (observed molecular mass 9.5 kDa), and it lacks the cysteine proteinase catalytic region, suggesting that it might act as a *trans*-inhibitor of cysteine proteinase(s)¹¹.

We performed experiments *in vitro* to assess the effect of Cer on the human liver cathepsins L and B and human papain. Cer was able to inhibit cathepsins L and B competitively ($K_i = 25$ nM and 30 nM, respectively; Supplementary Fig. 2b), but it could not inhibit papain even at 1 μ M (data not shown). These K_i values are similar to that described for other cathepsin inhibitors such as CTLA-2 β , a mammalian inhibitor of cathepsin L¹². Although the target(s) of Cer *in vivo* remain(s) to be identified, a *Drosophila* protein–protein interactions map has recently been generated showing that Cer interacts strongly with three proteins¹³ that are all cathepsins. Two of them correspond to cathepsin B group and seem to be expressed in the adult brain (as seen by northern blot assay; D.C., L. Zinck and

T.P., unpublished data). Those two cathepsins B constitutively lack the pro-inhibitor region. Cer is therefore probably needed to control their activity *in trans*.

To rescue the *cer^P* LTM phenotype, we introduced a genomic construction carrying the *cer⁺* gene with 1.5 kilobases (kb) of upstream sequence and 250 bp of downstream sequence (Fig. 2a). Rescue of the *cer^P* phenotype was observed in the presence of a single copy of the *cer⁺* transgene, indicating that this construct carries *cer* regulating sequences (Fig. 2d). In contrast, the presence of two copies of *cer⁺* could not rescue the *cer^P* LTM defect. Moreover, *cer⁺* flies expressing two additional genomic copies showed a decrease in LTM (Fig. 2d) but their STM remained normal (data not shown). Thus, *cer* overexpression affects LTM as severely as does a constitutive decrease in the concentration of *cer* mRNA. Taken together, these results show that the concentration of Cer must be within a short range of that in the wild type to form a normal LTM.

In which brain cells is *cer* specifically expressed? When the *cer* Gal4 enhancer-trap line was crossed with the reporter strain *P(UAS-mCD8-GFP)*, MBs were found to be strongly labelled together with large cells (Fig. 3a). Unfortunately, our anti-Cer antibody was not able to detect the peptide on tissues. Therefore, to identify directly the sites at which *cer* is expressed, we inserted the green fluorescent protein (GFP) open reading frame before the *cer* stop codon in the previously described genomic construct (Fig. 2a). Because a fragment carrying the same genomic sequences could rescue the LTM *cer* phenotype, we postulated that this fusion protein would be expressed in cells required for normal Cer function. Expression was observed in the MBs, and γ lobes were more strongly labelled than α/β lobes (Fig. 3b). MB neuron cell bodies were not strongly labelled. In the cortex and in the neuropile, large cells expressing Cer–GFP were found around the MB calyces and near MB lobes. These cells, which were also detected in *cer^P/P(UAS-mCD8-GFP)* individuals, had no axonal projections, indicating that they might be glial cells. Indeed, they did not express the nuclear neuronal cell marker Elav (ref. 14) (data not shown). In contrast, some of these cells were found to express the nuclear glial cell marker Repo (Fig. 3c–e). Some *cer*-expressing cells seem Repo-negative and Elav-negative. Because Elav is expressed in all neurons¹⁴, whereas the anti-Repo antibody does not label all glial cells, the large Cer-expressing cells most probably correspond to glia. Cer is therefore expressed in both MB neurons and adjacent glial cells.

To further delimit the brain structures in which the concentration of Cer is critical for LTM formation, we overexpressed *cer* with the

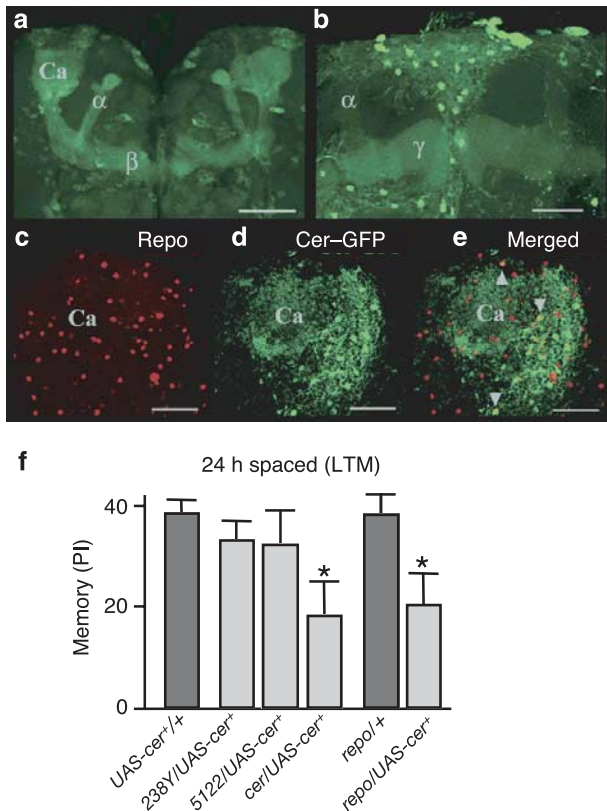


Figure 3 *cer* is expressed in MB neurons and in glial cells. **a**, Expression in a *cer^P–Gal4/P(UAS-mCD8-GFP)* line. **b**, Brain labelling of the 191 transgenic line carrying the Cer–GFP fusion protein, as viewed by confocal microscopy. **c**, A 191 brain stained with an antibody against Repo. **d**, Cer–GFP labelling. **e**, Colocalization of Cer–GFP and Repo. Regions of overlap are indicated by arrow heads. Scale bar, 40 μ m. **f**, Overexpression of *cer* with the *cer^P–Gal4* driver and with the glial driver *Repo–Gal4* induces an abnormal 24-h LTM. No effect was observed for overexpression in two different MB Gal4 lines, 238Y and 5122. Ca, calyx; α , alpha vertical lobe; β , beta medial lobe; γ , gamma medial lobe. PI, performance index. Results are means and s.e.m.; $n = 6$ –10 groups. Asterisk indicates significant differences in a *t*-test ($P < 0.05$).

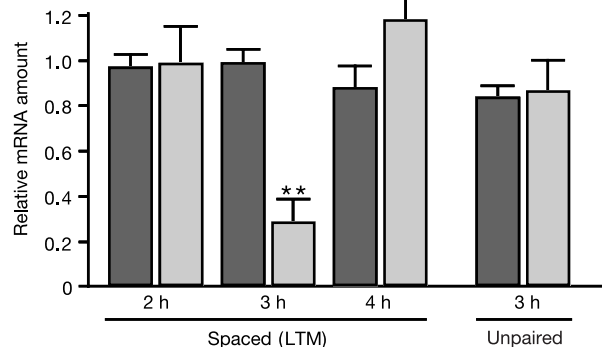


Figure 4 The level of *cer* mRNA is regulated after LTM conditioning. The level of *cer* mRNA is downregulated 3 h after LTM conditioning in the *CS* wild-type strain. An unpaired training does not affect the level of *cer* mRNA. Dark grey bars, untrained; pale grey bars, trained. Results are means and s.e.m. Asterisks indicate significant differences in a *t*-test ($P < 0.01$).

Gal4/UAS system. We observed a significant decrease in LTM conferred by the *cer*-Gal4 driver (Fig. 3f), confirming the dosage effect found with the genomic insert. Interestingly, a similar LTM defect was observed when *cer* was overexpressed with the *repo*-Gal4 glial-specific driver¹⁵. In contrast, no effect was seen when *cer* was overexpressed with the MB drivers 238Y (ref. 16) and 5122 (Fig. 3f) (the 5122 expression pattern is shown in Supplementary Fig. 3). These results indicate that the glial cells expressing *cer* might be involved in LTM formation. Indeed, neuron-like roles have recently been found for vertebrate glial cells. Some glial cells are able to integrate neuronal inputs, modulate synaptic activity, and process signals related to learning and memory¹⁷. *Cer* might be involved in the induction of neuronal plasticity by modulating crosstalk between neurons and glial cells.

Because the concentration of *Cer* seems to be a key factor in the establishment of LTM, we asked whether the *cer* transcript is regulated in the wild-type strain after LTM conditioning. We performed quantitative RT-PCR experiments at different time points after LTM training and found that the *cer* mRNA level decreased specifically 3 h after the end of training, in comparison with its level in untrained flies (Fig. 4). However, no significant change was observed 2 or 4 h after training (Fig. 4), indicating that the decrease in *cer* expression must occur in a narrow window of time to permit LTM formation. Similar results were obtained in western blot analysis: a 30% decrease in *Cer* level was observed between 3 and 4 h after conditioning (data not shown). Thus, the protein decay occurs slightly after the mRNA decay. Expression of *scs-fp* mRNA, the *cer*-adjacent gene, was not affected by LTM training (data not shown). Moreover, wild-type flies subjected to unpaired and repetitive odours and shocks, a regimen that does not induce learning, showed no variation in *cer* expression (Fig. 4). The sharp modulation of *cer* mRNA is therefore correlated with LTM formation.

What is the physiological meaning of the decrease in *cer* expression 3 h after LTM training? In the mouse, two waves of hippocampal mRNA synthesis are required for LTM formation after contextual fear conditioning¹⁸. Similar waves of gene expression have been described in *Hermissenda crassicornis* after associative conditioning¹⁹. Moreover, the simultaneous inhibition of multiple caspases (a family of cysteine proteinases) in the hippocampus blocks long-term, but not short-term, spatial memory²⁰. Our results extend these observations: we show that the expression of *cer*, which encodes an inhibitor of cysteine proteinases, is reduced for a short interval 3 h after training, thus probably leading to a transient activation of its cysteine proteinase(s) target(s). Significantly, prolonged cysteine proteinase activation is associated with neuronal degeneration in Alzheimer's disease: regions that contain maximal amounts of amyloid precursor protein also express maximal concentrations of cathepsin B and cathepsin L mRNA, indicating that the memory deficit of these patients might be linked to the deregulation of biochemical pathways involved in neuronal brain plasticity^{4,21}. We speculate that the memory deficit observed in flies with constitutively decreased concentrations of *Cer* might parallel this situation. □

Methods

Conditioning

The wild-type reference stock was *Canton-Special* (CS). The *cer*^P and all other strains used for memory experiments were outcrossed to flies of the CS background. Flies were conditioned by exposure to an odour paired with electric shocks and subsequent exposure to a second odour in the absence of shock, as described previously⁶.

Excisions and PCR rescue

Excision experiments were performed as described previously²². Genomic DNA adjacent to the P-element insertion was isolated by inverse PCR as described in <http://www.fruitfly.org/about/methods/inverse.pcr.html>.

Quantitative RT-PCR

Total RNA from the heads of conditioned or naive flies (1–2 µg) was used as a template for

reverse transcription using an oligo(dT) and SuperScript II reverse transcriptase (Invitrogen). Quantitative PCR was performed in accordance with the protocols provided with the LightCycler thermocycler (Roche) and the LightCycler – FastStart DNA Master SYBR Green I kit. The primers used were designed to amplify 60–80-bp fragments at the exon–intron boundary. Results were analysed in accordance with the instructions of the LightCycler software (version 3.5, Roche Molecular Biochemicals). The levels of *cer* and *scs-fp* mRNA were also normalized to the level of α -Tub84B mRNA.

Antibody production and western blot analysis

The *Cer*-glutathione S-transferase (*Cer*-GST) fusion protein was produced as described previously²³. Rabbits were injected dorsally several times every month for 3 months with 250 µg of the *Cer*-GST fusion protein mixed with Freund's adjuvant.

SDS-PAGE (with 70 µg of total adult head protein extract) and western blot analysis were performed as described previously²⁴. The membrane was incubated with polyclonal antiserum against the *Cer*-GST fusion protein (dilution 1:4,000) or with antiserum against Ciboulot (dilution 1:8,000), as described previously²³. The intensity of *Cer*-specific signals was normalized to that of Ciboulot-specific signals.

Cer inhibition assay

Cer activity was assayed as described for *Bombyx* cysteine proteinase inhibitor (BCPI) or CTLA-2 α (refs 12, 25), with a Safas flx spectrofluorimeter, *Cer* (0–200 nM), benzoyloxycarbonyl-Phe-Arg-7-amino-4-methylcoumarin as a substrate (3, 4.5 and 6 µM; Calbiochem, La Jolla, California), human liver cathepsin B and L (25 and 10 nM, respectively; Calbiochem) and papain (100 nM; Sigma, St Louis, Missouri).

Constructions

To create the genomic *cer*⁺ construct the clone BACR21D20 was digested with *Pst*I and *Xba*I. A 2,296-bp fragment (carrying the *cer*⁺ gene with 1.5 kb of sequence upstream of the transcription initiation site and 250 bp of downstream sequence) was cloned into pCaSpeR-4. Two transgenic lines, *GENX* and *GEND*, were obtained, with the P-element inserted in the second and third chromosome, respectively. Western blots were quantified and we observed that homozygous *GENX* and *GEND* had respectively twofold and threefold more *Cer* than CS (data not shown).

The *Cer*-GFP fusion protein construct was generated with the pCaSpeR-4 construction described above, by inserting GFP in phase before the *cer* stop codon. The *cer* 5' region until before the stop codon (*Bst*EII/*Bam*HI), the GFP open reading frame (*Bam*HI/*Xba*I) and the rest of *cer* (the 3' region from the stop codon and 250 bp of its downstream sequence (*Xba*I/*Pst*I) were amplified by PCR by using specific primers with the indicated restriction sites. All three PCR products were digested with appropriate restriction enzymes and the three products were cloned into *Bst*EII/*Pst*I-digested pCaSpeR-4 in a four-way ligation reaction.

To create the UAS-*cer*⁺ construct a full-length *cer*⁺ cDNA was isolated from the BDGP clone LP06209 and cloned into *Bgl*II/*Xba*I-digested pCaSpeR-UAS.

Immunohistochemical analysis

Drosophila brains were dissected as described previously⁶, then stained with the anti-Repo monoclonal antibody (dilution 1:100) or with the anti-Elav monoclonal antibody (dilution 1:10,000). Goat anti-mouse secondary antibody conjugated with Alexa-Fluor 568 (Interchim) was used at a dilution of 1:1,000. Expression was examined with a Leica (Wetzlar, Germany) TCS SP2 laser scanning confocal microscope.

Received 13 January; accepted 3 June 2004; doi:10.1038/nature02726.

- Bailey, C. H., Bartsch, D. & Kandel, E. R. Toward a molecular definition of long-term memory storage. *Proc. Natl Acad. Sci. USA* **93**, 13445–13452 (1996).
- Tully, T., Preat, T., Boynton, S. C. & Del Vecchio, M. Genetic dissection of consolidated memory in *Drosophila*. *Cell* **79**, 35–47 (1994).
- Alberini, C. M., Ghirardi, M., Metz, R. & Kandel, E. R. C/EBP is an immediate-early gene required for the consolidation of long-term facilitation in *Aplysia*. *Cell* **76**, 1099–1114 (1994).
- Callahan, L. M., Chow, N., Cheetham, J. E., Cox, C. & Coleman, P. D. Analysis of message expression in single neurons of Alzheimer's disease brain. *Neurobiol. Aging* **19**, S99–105 (1998).
- Waddell, S. & Quinn, W. G. Flies, genes, and learning. *Annu. Rev. Neurosci.* **24**, 1283–1309 (2001).
- Pascual, A. & Preat, T. Localization of long-term memory within the *Drosophila* mushroom body. *Science* **294**, 1115–1117 (2001).
- Isabel, G., Pascual, A. & Preat, T. Exclusive consolidated memory phases in *Drosophila*. *Science* **304**, 1024–1027 (2004).
- Dubnau, J. et al. The staufen/pumilio pathway is involved in *Drosophila* long-term memory. *Curr. Biol.* **13**, 286–296 (2003).
- Ito, K., Urban, J. & Technau, G. M. Distribution, classification and development of *Drosophila* glial cells in the late embryonic and early larval ventral nerve cord. *Wilhelm Roux Arch. Dev. Biol.* **204**, 284–307 (1995).
- Ito, K., Awano, W., Suzuki, K., Hiromi, Y. & Yamamoto, D. The *Drosophila* mushroom body is a quadruple structure of clonal units each of which contains a virtually identical set of neurones and glial cells. *Development* **124**, 761–771 (1997).
- Wiederanders, B. The function of propeptide domains of cysteine proteinases. *Adv. Exp. Med. Biol.* **477**, 261–270 (2000).
- Delaria, K. et al. Inhibition of cathepsin L-like cysteine proteases by cytotoxic T-lymphocyte antigen-2 beta. *J. Biol. Chem.* **269**, 25172–25177 (1994).
- Giot, L. et al. A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
- O'Neill, E. M., Rebay, I., Tjian, R. & Rubin, G. M. The activities of two Ets-related transcription factors required for *Drosophila* eye development are modulated by the Ras/MAPK pathway. *Cell* **78**, 137–147 (1994).
- Sepp, K. J., Schulte, J. & Auld, V. J. Peripheral glia direct axon guidance across the CNS/PNS transition zone. *Dev. Biol.* **238**, 47–63 (2001).

16. Armstrong, J. D., de Belle, J. S., Wang, Z. & Kaiser, K. Metamorphosis of the mushroom bodies; large-scale rearrangements of the neural substrates for associative learning and memory in *Drosophila*. *Learn. Mem.* **5**, 102–114 (1998).
17. Kurosinski, P. & Gotz, J. Glial cells under physiologic and pathologic conditions. *Arch. Neurol.* **59**, 1524–1528 (2002).
18. Igaz, L. M., Vianna, M. R., Medina, J. H. & Izquierdo, I. Two time periods of hippocampal mRNA synthesis are required for memory consolidation of fear-motivated learning. *J. Neurosci.* **22**, 6781–6789 (2002).
19. Epstein, H. T., Child, F. M., Kuzirian, A. M. & Alkon, D. L. Time windows for effects of protein synthesis inhibitors on Pavlovian conditioning in *Hermessenda*: behavioral aspects. *Neurobiol. Learn. Mem.* **79**, 127–131 (2003).
20. Dash, P. K., Blum, S. & Moore, A. N. Caspase activity plays an essential role in long-term memory. *Neuroreport* **11**, 2811–2816 (2000).
21. Maccioni, R. B., Munoz, J. P. & Barbeito, L. The molecular bases of Alzheimer's disease and other neurodegenerative disorders. *Arch. Med. Res.* **32**, 367–381 (2001).
22. Boquet, L., Hitier, R., Dumas, M., Chaminade, M. & Preat, T. Central brain postembryonic development in *Drosophila*: implication of genes expressed at the interhemispheric junction. *J. Neurobiol.* **42**, 33–48 (2000).
23. Boquet, L., Boujmaa, R., Carlier, M. F. & Preat, T. CIBOLUT regulates actin assembly during *Drosophila* brain metamorphosis. *Cell* **102**, 797–808 (2000).
24. Comas, D., Piulachs, M. D. & Belles, X. Fast induction of vitellogenin gene expression by juvenile hormone III in the cockroach *Blattella germanica* (L.) (Dictyoptera, Blattellidae). *Insect Biochem. Mol. Biol.* **29**, 821–827 (1999).
25. Kurata, M. *et al.* Bombyx cysteine proteinase inhibitor (BCPI) homologous to propeptide regions of cysteine proteinases is a strong, selective inhibitor of cathepsin L-like cysteine proteinases. *J. Biochem. (Tokyo)* **130**, 857–863 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M.F. Carlier, M. Hertzog and P. Mehlen for advice with biochemistry experiments; S. Brown and C. Talbot for confocal microscopy expertise; and T. Aoki, G. Didelot, G. Isabel and A. Pascual for their comments on this manuscript. This study was supported by a Human Frontier Science Program grant. E.P. was supported by the Ministère de la Recherche and the Fondation des Treilles. D.C. was supported by the Association pour la Recherche sur le Cancer and the European Community Marie Curie Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.P. (preat@iaf.cnrs-gif.fr).

Transmission of cutaneous leishmaniasis by sand flies is enhanced by regurgitation of fPPG

Matthew E. Rogers¹, Thomas Ilg^{2*}, Andrei V. Nikolaev³, Michael A. J. Ferguson³ & Paul A. Bates¹

¹Liverpool School of Tropical Medicine, University of Liverpool, Pembroke Place, Liverpool L3 5QA, UK

²Max-Planck-Institut für Biologie, Abteilung Membranbiochemie, Corrensstrasse 38, 72076 Tübingen, Germany

³Division of Biological Chemistry and Molecular Microbiology, School of Life Sciences, The Wellcome Trust Biocentre, University of Dundee, Dundee DD1 5EH, UK

* Present address: Intervet Innovation GmbH, Zur Propstei, 55270 Schwabenheim, Germany

Sand flies are the exclusive vectors of the protozoan parasite *Leishmania*¹, but the mechanism of transmission by fly bite has not been determined nor incorporated into experimental models of infection. In sand flies with mature *Leishmania* infections the anterior midgut is blocked by a gel of parasite origin, the promastigote secretory gel^{2,3}. Here we analyse the inocula from *Leishmania mexicana*-infected *Lutzomyia longipalpis* sand flies. Analysis revealed the size of the infectious dose, the underlying mechanism of parasite delivery by regurgitation, and the novel contribution made to infection by filamentous proteophosphoglycan (fPPG), a component of promastigote secretory gel found

to accompany the parasites during transmission. Collectively these results have important implications for understanding the relationship between the parasite and its vector, the pathology of cutaneous leishmaniasis in humans and also the development of effective vaccines and drugs. These findings emphasize that to fully understand transmission of vector-borne diseases the interaction between the parasite, its vector and the mammalian host must be considered together.

Leishmaniasis is a parasitic disease that now infects some 12 million people worldwide, causing severe morbidity and mortality⁴. Infection is initiated by distinct life cycle stages, metacyclic promastigotes, that are introduced into the skin by fly bite along with sand fly saliva^{5–7}. *Leishmania* are known to express various 'virulence factors' in the sand fly, which may facilitate transmission to, and infection of, the mammalian host^{8–12}. However, despite these discoveries our knowledge of parasite molecules that facilitate sand fly transmission is still limited. Furthermore, a number of key issues of transmission remain unresolved, such as the true infective dose, the mechanism of parasite delivery and the biological consequences of these upon infection. Importantly, in all *Leishmania*-vector combinations examined so far, a gel-like plug—the parasite-derived promastigote secretory gel (PSG)—blocks the anterior parts of the midgut coincident with the accumulation of metacyclic promastigotes^{2,3}. An important structural component of PSG is filamentous proteophosphoglycan (fPPG), an unusual mucin-like glycoprotein unique to *Leishmania*^{13,14}. Here we address these issues regarding transmission and reveal a novel contribution made by *L. mexicana* PSG to the infection process.

To begin to understand the nature of the infective inoculum, we determined the number and composition of *L. mexicana* parasites delivered during transmission (see Methods). We adapted a membrane feeding system to collect parasites egested by infected sand flies, revealing an average of 1,086 parasites delivered per bite, highly enriched in metacyclic promastigotes (86–98%) (Table 1). The only previous investigations quantifying egested parasites have been made using microcapillary forced feeding^{15,16}. When this method was employed on *L. mexicana*-infected sand flies an average of 105 promastigotes was collected per sand fly ($n = 19$), significantly ($P < 0.005$) underestimating the size of the average infective dose egested under voluntary feeding conditions. To address the origin of the average 10^3 parasites delivered during transmission, we quantified the foregut populations in sand flies. The average foregut population in pre-fed flies was 118 promastigotes ($n = 10$), and in post-fed flies was 53 promastigotes ($n = 13$) per sand fly. These populations are considerably smaller than the number of promastigotes that were actually egested. Therefore, the main source of egested parasites (>90%) lies behind the pharynx, thus clearly demonstrating active regurgitation of parasites from the oesophagus and behind, and not merely inoculation of the foregut population.

Next we investigated the efficiency of sand fly transmission by comparing infections caused by a single fly bite with those generated by a syringe inoculation. From the above results, a syringe inoculum of 10^3 metacyclic promastigotes was used to simulate natural transmission, and BALB/c and CBA/Ca mice expressed non-healing and healing phenotypes, respectively (Fig. 1a, b). However, when individual mice were infected by single fly bites different outcomes were quantitatively (BALB/c) and qualitatively (CBA/Ca) observed. BALB/c mice exhibited greatly accelerated lesion development, these appearing four to five weeks earlier than syringe inoculation and rapidly increasing in size without healing (Fig. 1a), whereas CBA/Ca mice developed non-healing chronic lesions (Fig. 1b). By analysing *Leishmania* infection using single fly bites we provide definitive evidence that delivery by sand flies makes a contribution to *Leishmania* infection beyond simple injection of parasites.

One plausible explanation for the results described above is that sand flies egest 'exacerbation factors' along with metacyclic pro-

Table 1 Egestion of metacyclic promastigotes by *L. longipalpis* infected with *L. mexicana*

Experiment	Size of total infection per fly ($\times 10^4$)	No. of metacyclic promastigotes per fly ($\times 10^4$)	No. of promastigotes egested per fly	Metacyclic promastigotes in egested population (%)	No. of metacyclic promastigotes egested per fly
1 ($n = 9$)	6.0 $\pm$ 0.9	1.62 (27%)	1,015 (1.7%)	98	995 (6.1%)
2 ($n = 18$)	3.6 $\pm$ 1.77	1.26 (35%)	703 (2.0%)	91	640 (5.1%)
3 ($n = 13$)	4.9 $\pm$ 1.0	1.52 (31%)	1,376 (2.8%)	86	1,183 (7.8%)
4 ($n = 16$)	7.5 $\pm$ 1.02	5.15 (69%)	1,251 (1.7%)	86	1,076 (2.1%)
Average	5.5	2.39 (43%)	1,086 (2.0%)	90	974 (4.1%)

Size of infections ($\pm$ s.e.m.) were determined from a pre-feed sample of ten flies in each experiment, and the percentage and number of metacyclic promastigotes were determined by morphological analysis. The number of promastigotes egested per fly is also expressed as a percentage of the total promastigote population. Similarly the number of metacyclic promastigotes egested per fly is also expressed as a percentage of the total number present in the pre-feed sample. Values of n for each independent experiment are the numbers of infected sand flies that delivered a single bite to the feeding apparatus.

mastigotes, which assist the establishment of the parasite in the mammalian host. Such factors could be parasite- and/or sand fly-derived. We exploited our method to collect metacyclic promastigotes from sand flies, and investigated whether any such exacerbation factors were co-egested along with the parasites. Egestion medium (with parasites removed) or control medium were mixed with 10^3 metacyclic promastigotes and inoculated into mice. BALB/c mice showed significantly enhanced lesion development and increased parasite burden when metacyclic promastigotes were co-inoculated with egestion medium (Fig. 1c). In CBA/Ca mice the infection was enhanced and the outcome was altered by egestion medium (Fig. 1d). The lesions quickly reached a large size, then decreased slightly after peaking but did not resolve, thus behaving like those produced by sand fly bites (Fig. 1b). Again there was a significant increase in parasite burden.

This led us to consider the identity of potential exacerbation factors. Previous work has shown that co-inoculation of parasites and sand fly salivary gland homogenate by syringe can exacerbate leishmaniasis^{6,7}, although the interpretation of these data in the context of natural transmission has been called into question^{1,17,18}. We confirmed, as expected, the presence of saliva in the egestion medium (Supplementary Fig. 1). We also considered the additional possibility that PSG is egested along with metacyclic promastigotes and may itself contribute to disease exacerbation. Egestion of PSG seems likely because we have shown here that metacyclic promastigotes are regurgitated from behind the pharynx, where PSG accumulates^{2,3}. We performed immunoblotting on the egestion medium of infected flies and this revealed the presence of fPPG (Fig. 2a), the only component of PSG detected in egestion medium. Analysis of sand fly gut homogenates showed that fPPG accumulated as the infections developed, reaching a peak on day 7 when flies were used in transmission experiments (Fig. 2b). The composition of PSG was further investigated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and immunoblotting (Fig. 2c). The main component of PSG was confirmed to be fPPG. We also detected other parasite-secreted glycans of lower molecular mass in PSG, together with a limited number of minor protein components. No lipophosphoglycan (LPG; 10–50 kDa) was detected in PSG or was free in the lumen of infected sand fly midguts, as predicted for a parasite surface glycoconjugate that mediates attachment to the midgut^{9,10} (Fig. 2c; Supplementary Fig. 2). We then infected sand flies with various *L. mexicana* null mutants selectively deficient in the following specific phosphoglycans: secreted acid phosphatase (sAP)¹⁹, PPG2²⁰ and LPG²¹. Analysis of PSG from these flies also indicated fPPG to be the dominant component (Fig. 2c). Mutants that cannot synthesize any phosphoglycans²² (*lpg2*^{-/-}) were unable to synthesize PSG or survive in sand flies. These data provide evidence for the existence of a second potential exacerbation factor, the parasite PSG, in addition to sand fly saliva.

Comparison of the disease-enhancing properties of saliva and PSG in BALB/c mice showed that saliva did not significantly exacerbate *L. mexicana* infection with 10^3 metacyclic promastigotes (Fig. 3a), and the final parasite burdens were actually lower than

controls. However, in CBA/Ca mice saliva did cause moderate disease exacerbation, but these lesions also resolved with time (Fig. 3b). In contrast, PSG caused substantial disease exacerbation in both BALB/c and CBA/Ca mice (Fig. 3c, d), and in the latter prevented healing of the lesions, which is characteristic of those generated by sand fly bite (Fig. 1b) and co-injection with egestion medium (Fig. 1d). In both mouse strains parasite burdens in PSG-co-inoculated mice were higher than in their respective controls. Because both saliva and PSG are delivered by sand flies during

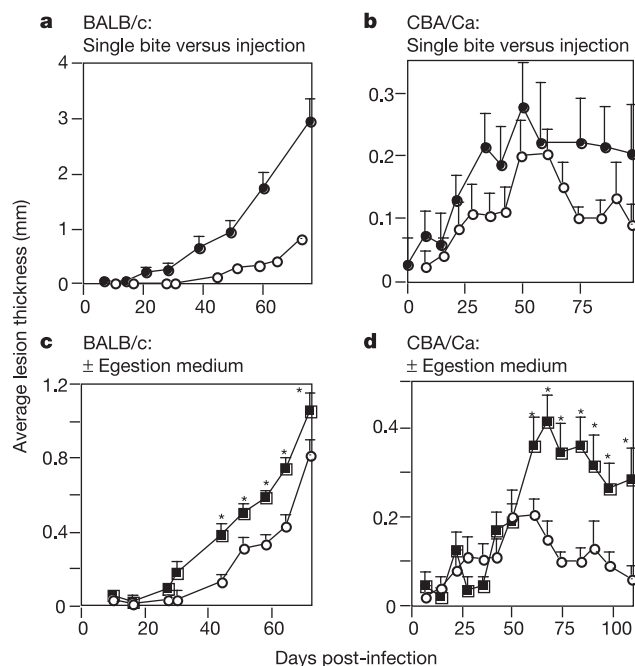


Figure 1 Sand flies make a significant contribution to transmission and egest exacerbation factors. **a, b**, Comparison of lesions developing from subcutaneous syringe injections of 10^3 sand-fly-egested metacyclic promastigotes (open circles) with those developing from single sand fly bites (filled circles), in BALB/c mice (**a**) and in CBA/Ca mice (**b**). Data for individual fly bite in BALB/c and CBA/Ca mice were derived from 16 and 10 mice, respectively. **c, d**, Comparison of footpad lesions developing from injection of 10^3 sand-fly-egested metacyclic promastigotes in 20 μ l medium alone (control, open circles) or in 20 μ l egestion medium (squares), in BALB/c mice (**c**) and in CBA/Ca mice (**d**). Lesion development was monitored by subtracting the width of the uninfected foot from the contralateral, infected foot, measured with a Vernier caliper. Final parasite loads in these mice were ($\pm$ s.e.m.) 1.71×10^8 (± 0.44), 2.43×10^8 (± 0.58) and 2.47×10^8 (± 0.65) for BALB/c mice infected by control injection, fly bite and co-inoculation with egestion medium, respectively, and 2.0×10^4 (± 1.12), 4.09×10^6 (± 0.62), and 5.31×10^6 (± 0.40) for corresponding groups of CBA/Ca mice. Asterisks represent statistically significant differences using an unpaired two-tailed Student's *t*-test ($P < 0.05$). Error bars represent 1 s.e.m.

transmission by bite, we also examined the effect of co-inoculation of both upon infection (Fig. 3e, f). In both BALB/c and CBA/Ca mice an intermediate course of lesion development resulted, indicating that the large disease exacerbation effect of PSG and the smaller effect of saliva, rather than acting synergistically, seemed to antagonize each other. Similarly, the parasite burdens in these mice were also intermediate. The antagonism between saliva and PSG presumably reflects differences in the underlying immunological mechanisms involved. These results show that PSG increased both the pathogenicity and survival of *L. mexicana*.

Next we considered the identity of the active ingredient(s) in PSG. First phosphoglycans (but not saliva) were adsorbed from

egestion medium by binding to DEAE-Sepharose, which consequently eliminated the exacerbatory properties of the egestion medium (Fig. 4a; Supplementary Fig. 1). Second, when PSG was ultracentrifuged the exacerbation effect segregated with the fPPG pellet fraction (Fig. 4b, c; Supplementary Fig. 3a). These data are consistent with the negative charge and high molecular weight of fPPG^{13,14}. Third, PSG was fractionated by SDS-PAGE, which showed fPPG retained by the stacking gel to be the active fraction (Fig. 4d, e; Supplementary Fig. 3b). Fourth, PSGs from *sap1/2*^{-/-}, *ppg2*^{-/-} and *lpg1*^{-/-} *L. mexicana* null mutants¹⁹⁻²¹ were co-inoculated with metacyclic promastigotes into CBA/Ca mice and were each found to exacerbate infections to the same extent as wild-type PSG (Supplementary Fig. 3d-f). By eliminating naturally occurring sAP and PPG2, or the possibility of contaminative LPG molecules within PSG, the only known molecule secreted by *L. mexicana* that possesses all of these properties, and the only molecule that we could detect in egestion medium is the fPPG component of PSG. From these data we cannot completely eliminate the existence of another bioactive component that makes a minor contribution to the exacerbation effect. However, based on

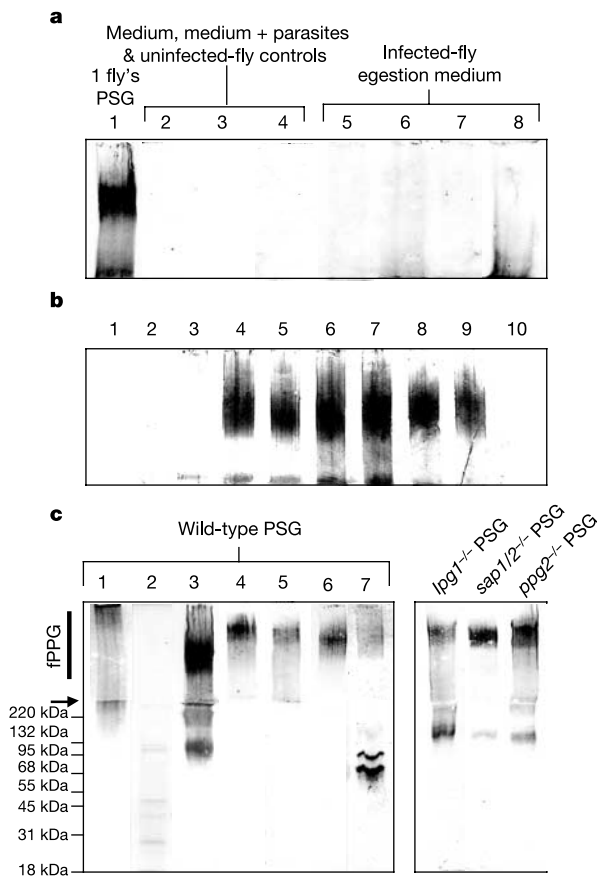


Figure 2 The fPPG component of *Leishmania* PSG is egested by infected sand flies. **a**, Stacking gel immunoblot probed with monoclonal antibody AP3 (ref. 27)—specific for phosphoglycan terminal mannose oligosaccharide caps—of PSG (lane 1), control medium (lane 2), control medium incubated with cultured metacyclic promastigotes (lane 3), egestion medium from uninfected sand flies (lane 4), and egestion medium from infected sand flies (lanes 5–8), corresponding to the four egestion experiments in Table 1, respectively. **b**, Stacking gel immunoblot probed with AP3 of infected gut homogenates from flies sampled on day 0 (lane 1) and thereafter at daily intervals (lanes 2–9), together with a blood-fed uninfected control sample (lane 10). Each lane represents the equivalent of one sand fly. **c**, Composition of wild-type PSG as revealed by Stains-All (lane 1), silver staining (lane 2) and immunoblots with monoclonal antibodies raised against *Leishmania* phosphoglycans. These antibodies include AP3 (lane 3), LT6 (lane 4) and LT15 (lane 5) (both specific for the galactose (Gal)-mannose (Man)-phosphate disaccharide repeat units), WIC108.3 (lane 6) (which recognizes both unsubstituted and substituted phosphosaccharides), and hydrofluoric acid dephosphorylated fPPG antisera (lane 7) (which specifically recognizes the polypeptide backbone of fPPG)^{13,27}. Each lane contained 1 μ g of PSG, except lane 2 which contained 5 μ g. Molecular mass markers are as indicated. The arrow indicates the interface between the stacking and resolving gels. The composition of PSG produced by *L. mexicana* mutants is as indicated, each having been probed with AP3.

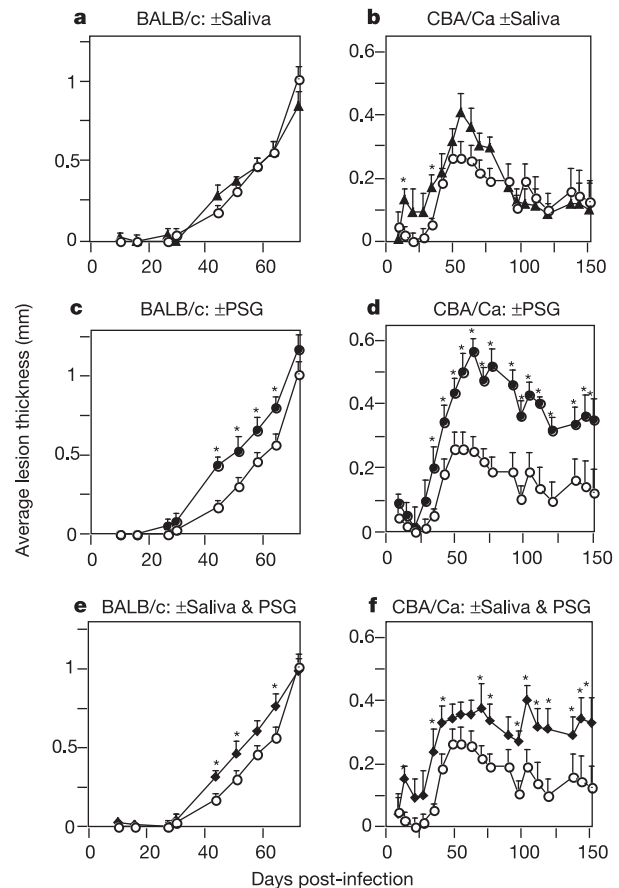


Figure 3 PSG enhances *L. mexicana* disease progression more than saliva. **a, c, e**, Comparison of lesions developing from injection of 10^3 cultured metacyclic promastigotes (control, open circles), +1 μ g saliva (**a**, triangles), +1 μ g PSG (**c**, filled circles), and +1 μ g saliva and 1 μ g PSG (**e**, diamonds) in BALB/c mice. **b, d, f**, An identical set of experiments, but performed in CBA/Ca mice. Lesions were monitored as before and upon termination processed for their total parasite loads: BALB/c control, $3.85 \times 10^8 (\pm 0.49 \text{ s.e.m.})$; +saliva, $1.81 \times 10^8 (\pm 0.39)$; +PSG, $4.44 \times 10^8 (\pm 0.39)$; +saliva and PSG $2.69 \times 10^8 (\pm 0.72)$; CBA/Ca control, $4.24 \times 10^5 (\pm 2.75)$; +saliva, $1.13 \times 10^4 (\pm 0.79)$; +PSG, $1.1 \times 10^6 (\pm 0.29)$; +saliva and PSG, $4.09 \times 10^5 (\pm 2.03)$. Asterisks represent statistically significant differences using an unpaired two-tailed *t*-test ($P < 0.05$). Error bars indicate 1 s.e.m.

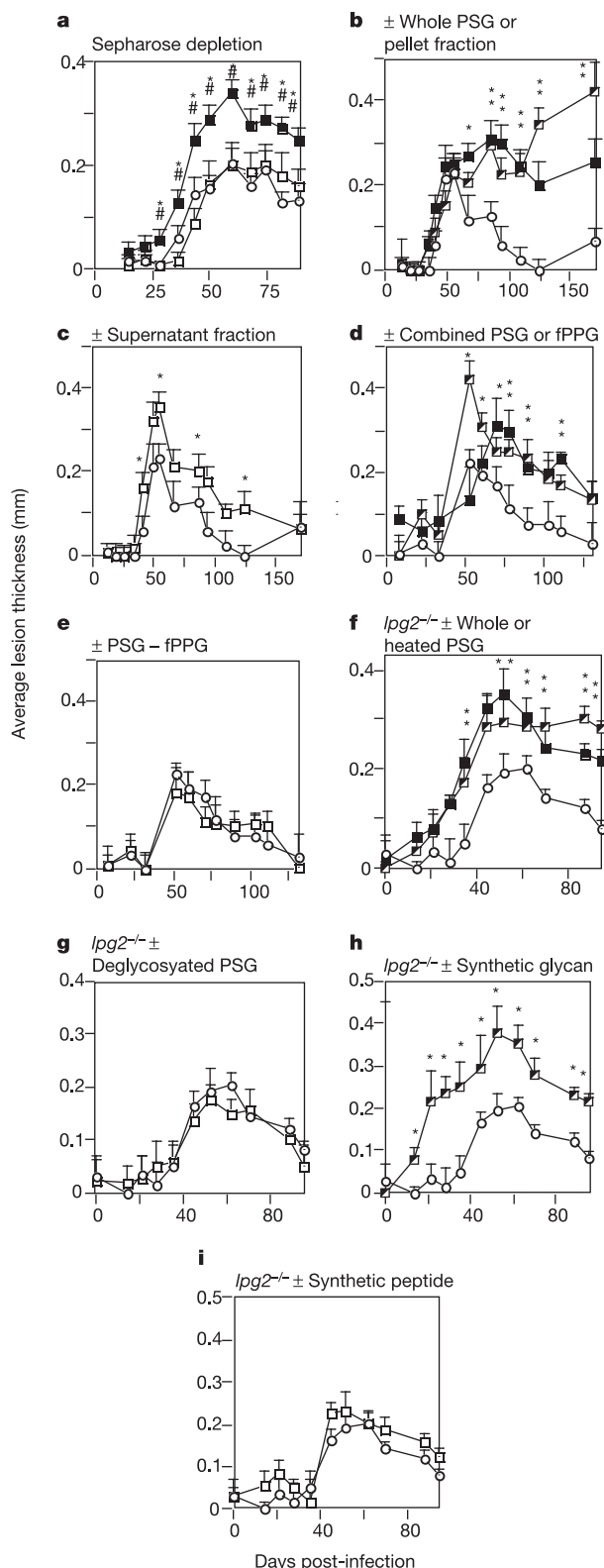


Figure 4 The fPPG fraction of PSG is responsible for enhancement of *Leishmania* infectivity. **a–i**, Comparison of lesion development in resistant CBA/Ca mice injected with 10^3 *L. mexicana* metacyclic promastigotes (controls, open circles) with mice co-injected with parasites and various components of PSG. Asterisks represent statistically significant differences compared with controls, using a two-tailed *t*-test ($P < 0.05$). **a**, Parasites resuspended in 20 μ l control medium, 20 μ l egestion medium (filled squares) or 20 μ l PSG-depleted egestion medium (open squares). Hash symbols indicate statistically significant differences between egestion medium and its depleted equivalent. **b, c**, After ultracentrifugation of PSG, parasites co-injected with whole PSG (filled squares), fPPG pellet (half-filled squares) (**b**) and supernatant fractions (open squares) (**c**). **d, e**, After SDS-PAGE separation, parasites co-injected with fPPG fraction (half-filled squares), all combined gel fractions (filled squares) (**d**) or the resolving gel components of PSG, that is, lacking fPPG (open squares) (**e**). Mild acid hydrolysis (pH 2, 60 °C, 1 h)²⁸ was used to selectively hydrolyse the acid labile phosphodiester linkages between the Gal-Man-phosphate repeat units of PSG. **f, g**, Using phosphoglycan-deficient *lpg2*^{-/-} *L. mexicana*²², parasites co-injected with PSG (filled squares), heat-treated PSG (half-filled squares) (**f**), or deglycosylated PSG (open squares) (**g**). **h, i**, Effect of synthetic phosphoglycan²³ (Gal(β 1–4)Man(α)–PO₃H.NH₃)₁₀–OH.NH₃ (half-filled squares) (**h**) and synthetic peptide (APSSSSAPSSSS) (open squares) (**i**), representing the disaccharide repeat core and the protein backbone characteristic of *Leishmania* phosphoglycans and fPPG, respectively^{13,14}. Final parasite loads reflected lesion development. Error bars indicate 1 s.e.m.

The glycans of fPPG were removed by mild acid hydrolysis and such deglycosylated fPPG completely lost its exacerbatory properties (Fig. 4f, g; Supplementary Fig. 3g–i). Furthermore, a chemically synthesized *Leishmania* phosphoglycan²³ proved to be highly effective at exacerbating *lpg2*^{-/-} *L. mexicana* infections (Fig. 4h; Supplementary Fig. 3j), whereas the serine-rich repeat motif, synthesized to mimic the backbone of *L. mexicana* fPPG^{13,14}, did not influence infections at all (Fig. 4i). Therefore, these data show that the glycan moieties of the egested fPPG were responsible for the exacerbation of infection.

Thus a new picture of the transmission of this vector-borne disease has emerged, showing how sand flies act as vectors of leishmaniasis. Proof of regurgitation as the mechanism of natural parasite delivery led to the discovery of a new component of the infective inoculum, the PSG. This can be added to the existing part that PSG plays in creating a ‘blocked fly’ that experiences difficulty in feeding, leading to multiple and longer feeding attempts and more opportunities for transmission^{2,3,24}. The active ingredient in PSG that led to long-term disease exacerbation was identified as fPPG, showing the advantage of secreted phosphoglycans for transmission. Thus this report demonstrates that *L. mexicana* uses secreted fPPG to manipulate both insect vector and mammalian host. This dual role of fPPG ensures efficient transmission and proves to be a highly beneficial survival strategy. Given the selection pressures in the evolution of infectious diseases, it is not surprising that *Leishmania* has found intriguing and novel ways to secure its own transmission. □

Methods

Female *L. longipalpis* were infected by feeding on 2×10^6 *L. mexicana* lesion amastigotes per ml rabbit blood³. *L. longipalpis* is not the natural vector of *L. mexicana* but has been shown to support the complete development of the parasite and transmit the infection to mice through biting³. For the generation of mutant PSG, flies were infected with 6×10^6 *sap1/2*^{-/-} (ref. 19), *ppg2*^{-/-} (ref. 20) and *lpg1*^{-/-} (ref. 21) amastigotes/promastigotes in their first passage, to retain their infectivity. This modification was required because these mutants did not infect flies as efficiently as wild-type parasites. Flies were maintained for 7–9 days to allow mature infections to develop. To obtain egested metacyclics, groups of 50 *L. mexicana*-infected flies then had the opportunity to feed through a chick skin on promastigote culture medium³. Individual flies that alighted on the apparatus and began feeding were observed, immediately removed from the cage when they retracted their mouthparts, and kept in a separate container. After 1 h the feeder was withdrawn and the egestion medium processed to reveal the number of egested parasites regurgitated by sand flies in 2 ml of egestion medium, which were concentrated by centrifugation and counted. The sand flies were dissected and morphological analysis of parasites was

the evidence, we conclude that the main active ingredient of PSG is fPPG.

The protein backbone of fPPG includes a repetitive serine-rich motif that is very heavily glycosylated^{13,14}. Therefore, lastly we investigated whether disease exacerbation was associated with the peptide and/or glycan moieties of fPPG. For these experiments we used wild-type and also phosphoglycan-deficient *lpg2*^{-/-} parasites²².

performed³. Foregut populations were regarded as parasites located anterior of the oesophagus–pharynx junction. Microcapillary forced feeding was performed as described²⁵.

Groups of 8–10 mice were infected either by individual sand fly bite or by subcutaneous inoculation of 20 µl containing metacyclic promastigotes, with or without test materials (such as PSG and saliva), in the upper surface of the right hind foot. Sand flies with 7-day-old infections were introduced singly into a cage with an anaesthetized mouse that was screened from the fly except for its right hind leg. Flies were removed from the cage when they had taken a single feeding attempt and were then checked for infection by microscopy. Lesion development was monitored by measuring the swelling of the foot with Vernier calipers. At the end of experiments mice were humanely killed, their feet removed and parasite burdens determined by homogenization and counting.

Cultured metacyclic promastigotes were generated as described²⁶, except that Grace's culture medium (Invitrogen), supplemented with 10% (v/v) fetal calf serum and adjusted to pH 5.5, was used. Such *L. mexicana* cultured metacyclic promastigotes were of equal infectivity to those egested by sand flies (Supplementary Information Fig. 4), and were used in experiments in which higher numbers of parasites were required.

High molecular weight PPGs were analysed by SDS–PAGE using an enlarged 4% polyacrylamide stacking gel, as described¹³, and by immunoblotting with monoclonal antibodies AP3, LT6, LT15, WIC108.3 and anti-HF-deglycosylated fPPG^{13,27}. Infected gut homogenates were prepared as described³, centrifuged to remove cells and debris, and the supernatant mixed with an equal volume of 2 × gel sample buffer. Proteophosphoglycans were concentrated by adsorption from 1 ml volumes of egestion medium by incubation with 20% (v/v) DEAE-Sephacrose (Pharmacia), beads were washed in PBS (10 mM sodium phosphate, 145 mM sodium chloride, pH 7.2) and proteophosphoglycans were eluted into 50 µl gel sample buffer.

Salivary glands were obtained by dissection of 6–8-day-old female flies in PBS on cooled slides and transferred to a clean slide. This was followed by puncture of individual glands with fine entomological needles in 5 µl PBS. The saliva released was centrifuged and stored at –70 °C. Crude whole salivary gland homogenate or sonicate was deliberately not used to avoid contamination of saliva with salivary gland epithelial cell components¹. Salivary gland homogenate was found to produce significantly more disease exacerbation than saliva (Supplementary Information Fig. 5). PSG plugs were isolated from the midguts of day 7 infected flies³, and PSG was separated from the parasites by dispersing in PBS (5 µl per plug) followed by four rounds of centrifugation (10,000 g, 5 min) and retention of supernatant, then stored at –70 °C. Saliva and PSG were quantified using the BCA (bicinchoninic acid) protein assay (Pierce). Sand flies contained, on average, 0.97 µg saliva and 0.86 µg PSG per fly.

Filamentous PPG was fractionated from PSG by non-reducing SDS–PAGE, followed by resuspension in culture medium. Twelve PSG plugs were separated (three plugs × four lanes) and the stacking gel (fPPG fraction) was removed from and processed separately to the resolving gel (other PPGs and protein fraction). Equal volumes of gel, including a control piece of acrylamide (control fraction), were washed in PBS then homogenized in 1 ml culture medium using a Teflon coated pestle until it could be injected freely using a 21-gauge needle. After centrifugation (10,000 g, 5 min) the supernatants were combined with parasites to infect mice. Mild acid hydrolysis (pH 2, 60 °C, 1 h) was performed as described²⁸. Synthetic phosphoglycan was synthesized as described²³ and the peptide mimicking the fPPG backbone was prepared commercially (Pepsyn). Each were used at 1 µg per infection.

Received 3 February; accepted 18 May 2004; doi:10.1038/nature02675.

1. Sacks, D. L. & Kamhawi, S. Molecular aspects of parasite-vector and vector-host interactions in leishmaniasis. *Annu. Rev. Microbiol.* **55**, 453–483 (2001).
2. Stierhof, Y.-D. *et al.* Filamentous proteophosphoglycan secreted by *Leishmania* promastigotes forms gel-like three-dimensional networks that obstruct the digestive tract of infected sandfly vectors. *Eur. J. Cell Biol.* **78**, 675–689 (1999).
3. Rogers, M. E., Chance, M. L. & Bates, P. A. The role of promastigote secretory gel in the origin and transmission of the infective stage of *Leishmania mexicana* by the sandfly *Lutzomyia longipalpis*. *Parasitol.* **124**, 495–507 (2002).
4. UNDP/World Bank/WHO. *Leishmaniasis* (<http://www.who.int/tdr/diseases/leish/diseaseinfo.htm>) (2004).
5. Sacks, D. L. & Perkins, P. V. Identification of an infective stage of *Leishmania* promastigotes. *Science* **223**, 1417–1419 (1984).
6. Titus, R. G. & Ribeiro, J. M. Salivary gland lysates from the sand fly *Lutzomyia longipalpis* enhance *Leishmania* infectivity. *Science* **239**, 1306–1308 (1988).
7. Belkaid, Y. *et al.* Development of a natural model of cutaneous leishmaniasis: powerful effects of vector saliva and saliva preexposure on the long-term outcome of *Leishmania major* infection in the mouse ear dermis. *J. Exp. Med.* **188**, 1941–1953 (1998).
8. Schleim, Y., Jacobson, R. L. & Messer, G. *Leishmania* infections damage the feeding mechanism of the sandfly vector and implement parasite transmission by bite. *Proc. Natl Acad. Sci. USA* **89**, 9944–9948 (1992).
9. Pimenta, P. F. *et al.* Stage-specific adhesion of *Leishmania* promastigotes to the sandfly midgut. *Science* **256**, 1812–1815 (1992).
10. Sacks, D. L. *et al.* The role of phosphoglycans in *Leishmania*-sand fly interactions. *Proc. Natl Acad. Sci. USA* **97**, 406–411 (2000).
11. Späth, G. F. *et al.* Lipophosphoglycan is a virulence factor distinct from related glycoconjugates in the protozoan parasite *Leishmania major*. *Proc. Natl Acad. Sci. USA* **97**, 9258–9263 (2000).
12. Späth, G. F. *et al.* Persistence without pathology in phosphoglycan-deficient *Leishmania major*. *Science* **301**, 1241–1243 (2003).
13. Ilg, T. *et al.* Purification and structural characterization of a filamentous, mucin-like proteophosphoglycan secreted by *Leishmania* parasites. *J. Biol. Chem.* **271**, 21583–21596 (1996).
14. Ilg, T. Proteophosphoglycans of *Leishmania*. *Trends Parasitol.* **16**, 489–497 (2000).
15. Warburg, A. & Schleim, Y. The effect of post-bloodmeal nutrition of *Phlebotomus papatasi* on the

- transmission of *Leishmania major*. *Am. J. Trop. Med. Hyg.* **35**, 926–930 (1986).
16. Saraiva, E. M. B. *et al.* Changes in lipophosphoglycan and gene expression associated with the development of *Leishmania major* in *Phlebotomus papatasi*. *Parasitol.* **111**, 275–287 (1995).
17. Kamhawi, S., Belkaid, Y., Modi, G., Rowton, E. & Sacks, D. L. Protection against cutaneous leishmaniasis resulting from bites of uninfected sand flies. *Science* **290**, 1351–1354 (2000).
18. Castro-Sousa, F. *et al.* Dissociation between vasodilation and *Leishmania*-enhancing effects of sand fly saliva and maxadilan. *Mem. Inst. Oswaldo Cruz* **96**, 997–999 (2001).
19. Wiese, M., Ilg, T., Lottspeich, F. & Overath, P. Ser/Thr-rich repetitive motifs as targets for phosphoglycan modifications in *Leishmania mexicana* secreted acid phosphatase. *EMBO J.* **14**, 1067–1074 (1995).
20. Gopfert, U. *et al.* Proteophosphoglycans of *Leishmania mexicana*. Molecular characterization of the *Leishmania mexicana* ppg2 gene encoding the proteophosphoglycans aPPG and pPPG2 that are secreted by amastigotes and promastigotes. *Biochem. J.* **344**, 787–795 (1999).
21. Ilg, T. Lipophosphoglycan is not required for infection of macrophages or mice by *Leishmania mexicana*. *EMBO J.* **19**, 1953–1962 (2000).
22. Ilg, T., Demar, M. & Harbecke, D. Phosphoglycan repeat-deficient *Leishmania mexicana* parasites remain infectious to macrophages and mice. *J. Biol. Chem.* **276**, 4988–4997 (2001).
23. Nikolaev, A. V., Chudek, J. A. & Ferguson, M. A. J. The chemical synthesis of *Leishmania donovani* phosphoglycan via polycondensation of a glycosyl hydrogen phosphonate monomer. *Carbohydr. Res.* **272**, 179–189 (1995).
24. Beach, R., Kiilu, G. & Leeuwenburg, J. Modification of sandfly biting behaviour by *Leishmania* leads to increased parasite transmission. *Am. J. Trop. Med. Hyg.* **34**, 278–282 (1985).
25. Hertig, M. & McConnell, E. Experimental infection of Panamanian *Phlebotomus* sandflies with *Leishmania*. *Exp. Parasitol.* **14**, 92–106 (1963).
26. Bates, P. A. & Tetley, L. *Leishmania mexicana*: induction of metacyclogenesis by cultivation of promastigotes at acidic pH. *Exp. Parasitol.* **76**, 412–423 (1993).
27. Ilg, T., Harbecke, D., Wiese, M. & Overath, P. Monoclonal antibodies directed against *Leishmania* secreted acid phosphatase and lipophosphoglycan. *Eur. J. Biochem.* **217**, 603–615 (1993).
28. Bates, P. A., Hermes, I. & Dwyer, D. M. Golgi-mediated post-translational processing of secretory acid phosphatase by *Leishmania donovani* promastigotes. *Mol. Biochem. Parasitol.* **39**, 247–256 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The technical assistance of D. Moor and J. Archer is acknowledged. We thank M. Hajmova and P. Volf for assistance with the forced feeding technique and M. Wiese and P. Overath for antibodies and *Leishmania* mutants. This work received financial support from the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR) and the Wellcome Trust, UK.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.A.B. (pbates@liv.ac.uk).

Reciprocal regulation of haem biosynthesis and the circadian clock in mammals

Krista Kaasik^{1,2} & Cheng Chi Lee^{1*}

¹Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

²Department of Biotechnology, Institute of Molecular and Cell Biology, Tartu University, Tartu, 51010, Estonia

* Present address: Department of Biochemistry and Molecular Biology, University of Texas Health Science Center, Houston, Texas 77030, USA

The circadian clock is the central timing system that controls numerous physiological processes. In mammals, one such process is haem biosynthesis, which the clock controls through regulation of the rate-limiting enzyme aminolevulinic synthase 1 (Alas1)^{1,2}. Several members of the core clock mechanism are PAS domain proteins, one of which, neuronal PAS 2 (NPAS2), has a haem-binding motif^{3,4}. Indeed, haem controls activity of the BMAL1–NPAS2 transcription complex *in vitro* by inhibiting DNA binding in response to carbon monoxide³. Here we show that haem differentially modulates expression of the mammalian *Period* genes *mPer1* and *mPer2* *in vivo* by a mechanism involving NPAS2 and mPER2. Further experiments show that mPER2

positively stimulates activity of the BMAL1–NPAS2 transcription complex and, in turn, NPAS2 transcriptionally regulates *Alas1*. Vitamin B12 and haem compete for binding to NPAS2 and mPER2, but they have opposite effects on *mPer2* and *mPer1* expression *in vivo*. Our data show that the circadian clock and haem biosynthesis are reciprocally regulated and suggest that porphyrin-containing molecules are potential targets for therapy of circadian disorders.

To investigate the role of haem in the circadian clock mechanism, we first observed that haem could synchronize clock gene expression in NIH 3T3 cells (Supplementary Fig. 1). Next, mice were injected intraperitoneally (i.p.) with haem or solvent during subjective day and night to determine the effect of haem on clock gene expression *in vivo*. Northern blot analysis indicated that haem, but not solvent, dampened the expression of *mPer2* but enhanced that of *mPer1*. The effects of haem on *mPer1* and *mPer2* expression were most prominent during subjective night. No apparent change in the expression of *Bmal1*, *Npas2* or *Cry1* was observed in wild-type mice (Fig. 1).

Wheel-running activity in mice showed that injection of haem during subjective day at circadian time (CT) 6 did not change locomotor behaviour as compared with injection of solvent (Supplementary Fig. 2a). By contrast, haem injected during subjective night (CT22) inhibited wheel-running activity, whereas activity was unchanged in mice that received solvent (Supplementary Fig. 2b). These behavioural and molecular observations suggest that the effects of haem are likely to be mediated by circadian-regulated factors.

To identify these circadian factors, *Npas2* mutant (*Npas2*^{m/m}), *mPer1*-null (*mPer1*^{-/-}) and *mPer2* mutant (*mPer2*^{m/m}) mice were injected i.p. with haem at various CTs. Northern blot analysis showed that the differential effects of haem on *mPer1* and *mPer2* were attenuated in *Npas2*^{m/m} and *mPer2*^{m/m} mice, but not in *mPer1*^{-/-} mice (Supplementary Fig. 3a,b). These observations indicate that several regulators, including NPAS2 and mPER2, may mediate the effects of haem on *mPer1* and *mPer2* expression.

We investigated the relationship between mPER2 function and *Npas2* expression *in vivo*. Northern blot analysis showed that *Npas2* was expressed between CT20 and CT4 in wild-type mice

(Figs 1 and 2a). In both *mPer2*^{m/m} and *mPer2*-null (*mPer2*^{-/-}) mice (data not shown), however, no detectable *Npas2* expression was observed at any CT analysed (Fig. 2a). Consistent with *in situ* hybridization studies³, *Clock* transcript levels were indistinguishable between *mPer2*^{m/m} and wild-type mice (Fig. 2a). In *mPer1*^{-/-} mice, expression of *Npas2* was decreased, suggesting that mPER1 has a lesser role than mPER2 in regulating *Npas2* expression (Supplementary Fig. 4a). Together, these observations suggest that the target of mPER2 is BMAL1–NPAS2, rather than the BMAL1–CLOCK transcription complex.

To test this possibility, mPER2 regulation of BMAL1–NPAS2 and BMAL1–CLOCK transcription complex activity was investigated by using luciferase reporter constructs based on *mPer1* and *mPer2* promoters. Both BMAL1–CLOCK and BMAL1–NPAS2 transcription complexes strongly activated *mPer1* and *mPer2* promoters, as previously reported^{6–9}. Coexpression of mPER2 further enhanced BMAL1–NPAS2 but not BMAL1–CLOCK transcription activity with both promoters (Fig. 2b and Supplementary Fig. 4b). Coexpression of CRY1 repressed the mPER2-induced BMAL1–NPAS2 transcriptional control of both promoters to levels below those observed for BMAL1–NPAS2-dependent activation. Expression of mPER1 enhanced BMAL1–NPAS2-dependent activation of the *mPer1* promoter but not the *mPer2* promoter (Fig. 2b and Supplementary Fig. 4b). These findings show that mPER2 is a positive stimulator of BMAL1–NPAS2 transcriptional activity and a chief regulator of NPAS2 *in vivo*.

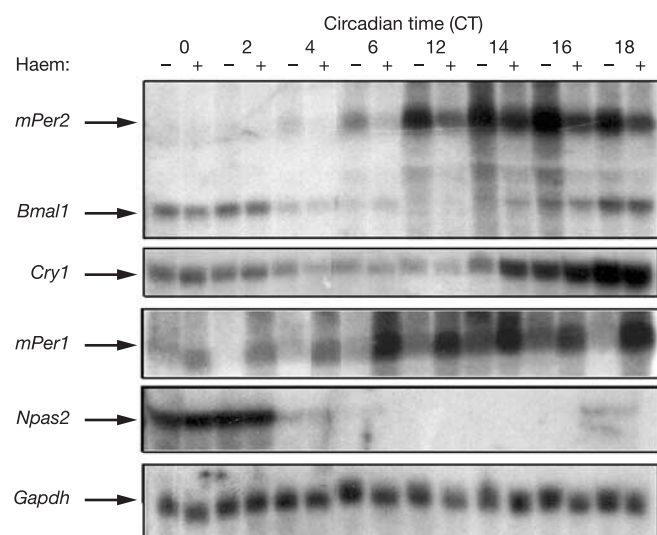


Figure 1 Haem modulates clock gene expression in liver. Expression of *mPer2*, *mPer1*, *Bmal1*, *Cry1* and *Npas2* messenger RNA was analysed by northern blotting with the respective cDNA probes. *Gapdh* mRNA was monitored as an internal control.

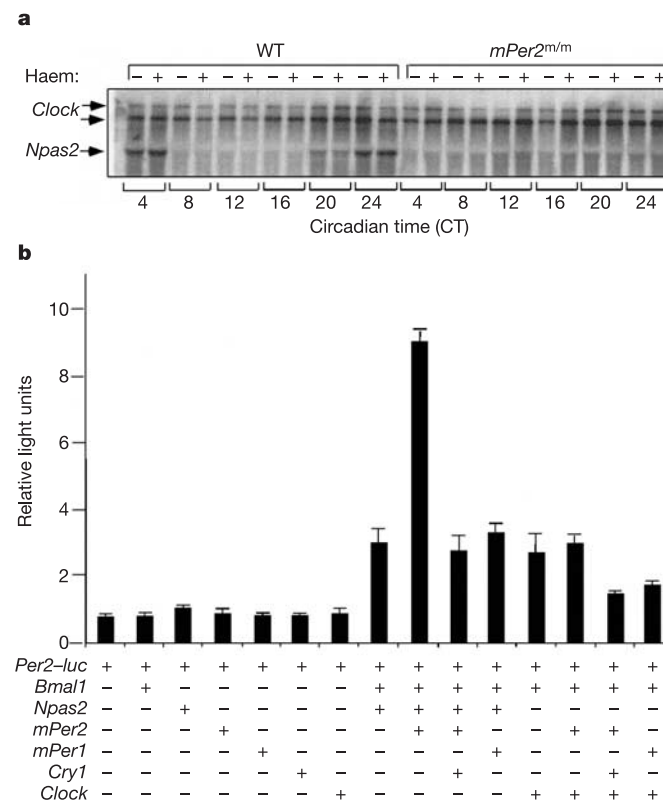


Figure 2 NPAS2 is regulated by mPER2, and BMAL1–NPAS2 is a molecular target of mPER2. **a**, Northern blot analysis using ³²P-labelled *Clock* and *Npas2* cDNAs as probes. RNA was isolated from wild-type and *mPer2*^{m/m} mice injected i.p. with solvent or haem. **b**, Reporter assay studies using the *mPer2* promoter. COS-1 cells were transfected with the reporter construct *mPer2-luc* and various combinations of expression constructs (Methods). Luciferase activity in samples transfected with reporter plasmid alone was arbitrarily set at 1.0. Error bars indicate the s.e.m. ($n = 6$).

Loss of both mPER1 and mPER2 function leads to deregulated *Alas1* expression¹, but the mechanism of their circadian regulation of *Alas1* expression is unclear. Because BMAL1 and NPAS2 are basic helix–loop–helix transcription factors, we tested whether they could regulate *Alas1* expression directly. Northern blot analysis showed that the robust expression of *Alas1* observed in wild-type mice was markedly dampened in *Npas2*^{m/m} mice (Fig. 3a). Unlike the robust inhibitory effects observed in wild-type mice, haem repression of *Alas1* expression was attenuated in *Npas2*^{m/m} mice except between CT12 and CT16. The weak circadian amplitude of *Alas1* at CT12 and CT16 suggested possible control by redundant clock transcription regulators or residual transcriptional activity of the mutant NPAS2 (ref. 10).

To confirm that the BMAL1–NPAS2 transcription complex is a direct regulator of the *Alas1* promoter, a 5.2-kilobase (kb) promoter fragment derived from the murine *Alas1* gene was cloned into the luciferase reporter vector pGL2-basic to generate an *Alas1*–*luc* construct. The presence of BMAL1–NPAS2 stimulated *Alas1*–*luc* promoter activity by about 2-fold (Fig. 3b). A 3.5-fold stimulation of promoter activity was observed when NPAS2, BMAL1 and mPER2 were added together. Individually, mPER2, BMAL1 or NPAS2 did not stimulate *Alas1*–*Luc* activity. These results show that NPAS2 is a principal transcriptional regulator of *Alas1* *in vivo*.

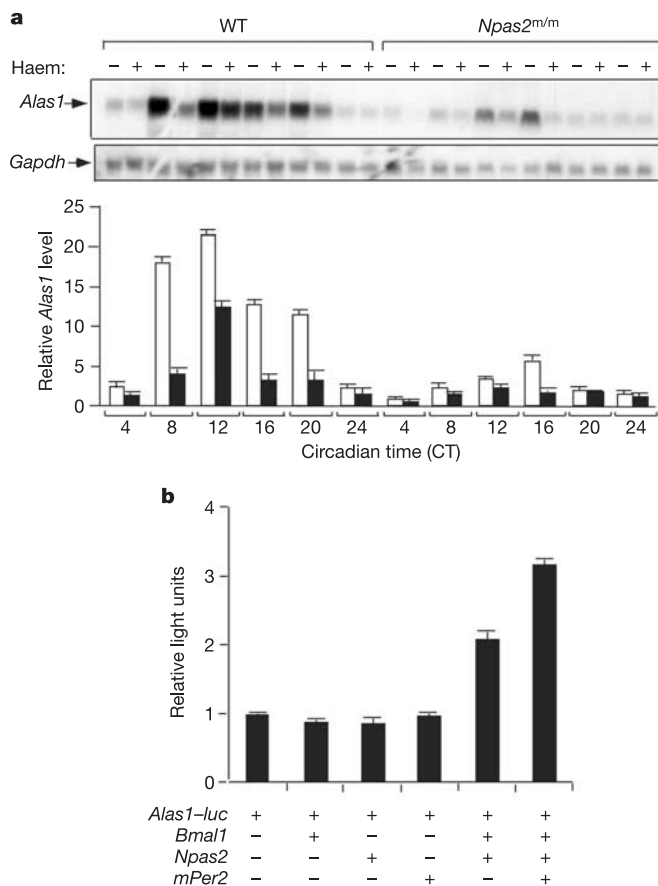


Figure 3 NPAS2 is a principal regulator of ALAS1. **a**, Northern blot analysis of RNA from wild-type and *Npas2*^{m/m} mice injected i.p. with solvent or haem. Blots were hybridized with ³²P-labelled *Alas1* and *Gapdh* cDNAs. The level of *Alas1* mRNA was normalized to that of *Gapdh* mRNA, and the lowest relative level was arbitrarily set at 1.0. Error bars indicate the s.d. (*n* = 2). **b**, Reporter assay studies with the *Alas1* promoter (Methods). Luciferase activity in cells transfected with the reporter plasmid *Alas1*–*luc* alone was arbitrarily set at 1.0. Error bars indicate the s.e.m. (*n* = 6).

The presence of a PAS motif¹¹ in both proteins raised the possibility that mPER2 modulation of NPAS2 activity might involve haem interaction. We used an assay based on haem–agarose affinity chromatography¹² to investigate haem binding by mPER2. To investigate the role of the PAS domain, we cloned an *mPer2*^{m/m} complementary DNA encoding an 87-amino-acid in-frame deletion of the PAS-B domain into the pCMV–Sport2 expression vector⁵. Proteins were produced by transfecting the expression constructs pCMV–Sport2–mPer2 and pCMV–Sport2–mPer2^{m/m} into COS-1 cells. For controls, mPER2 binding to unconjugated agarose and GAPDH binding to a haem–agarose matrix were also monitored.

Western blot analysis showed that wild-type mPER2 binds efficiently to the haem–agarose matrix (Fig. 4a). COS-1 cells not transfected with the *mPer2* expression vector showed no detectable mPER2. GAPDH did not bind haem–agarose, and unconjugated agarose matrix did not bind mPER2 (Fig. 4a). Both *mPer2* and *mPer2*^{m/m} expression constructs produced comparable amounts of proteins in COS-1 cells (Fig. 4b); however, mutant mPER2 bound

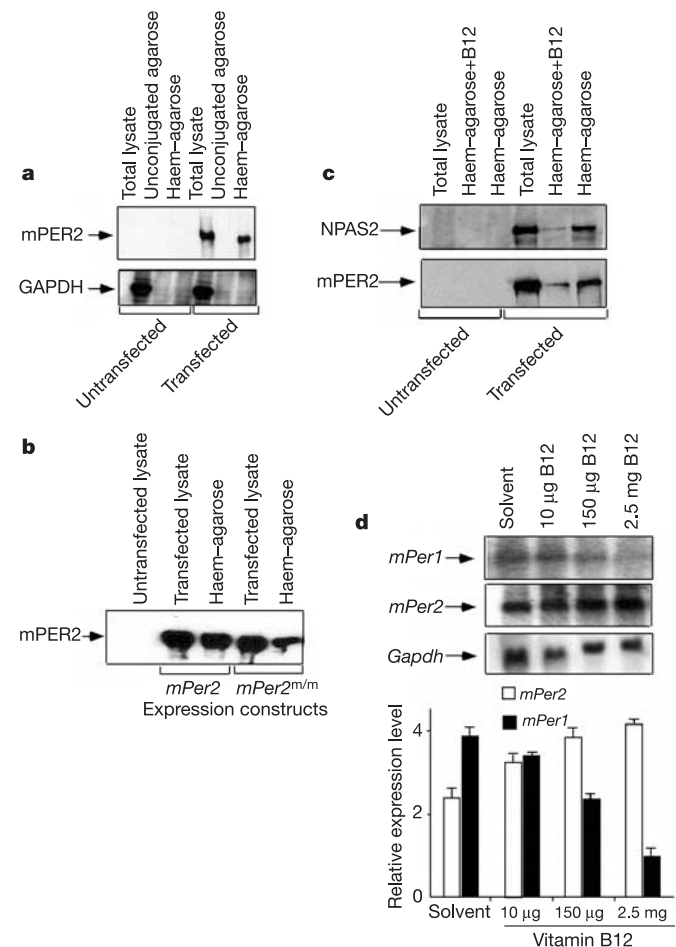


Figure 4 Haem and vitamin B12 binding by mPER2 and NPAS2. **a**, Western blot analysis of mPER2 protein in samples affinity purified by haem–conjugated or unconjugated agarose chromatography. **b**, Western blot analysis of wild-type and mutant mPER2 isolated by haem–agarose, using antisera against mPER2. **c**, Western blot analysis of NPAS2 and mPER2 isolated by haem–agarose in the presence of vitamin B12, using monoclonal antibodies against NPAS2 and antisera against mPER2. **d**, Northern blot analysis of *mPer1* and *mPer2* mRNA in livers 4 h after mice were injected with vitamin B12 or solvent. Bottom panel shows the expression of *mPer1* and *mPer2* relative to that of *Gapdh*. Error bars indicate the s.d. (*n* = 2).

the haem-agarose matrix with much lower efficiency than did wild-type protein. Further deletion of the PAS-A domain resulted in the production of much less mutant protein, suggesting protein instability (data not shown).

Studies have shown that cobalamin can modulate human and rodent circadian rhythm, but the underlying mechanism is unclear^{13,14}. Given the similarity of the porphyrin ring structure of haem and cobalamin, we tested whether a vitamin B12 analogue, cyanocobalamin, could inhibit haem binding to NPAS2 or mPER2. Binding of NPAS2 and mPER2 to the haem-agarose matrix was greatly decreased when an equal molarity of cyanocobalamin was added during the binding step (Fig. 4c). These findings suggest that cobalamin can compete with haem for PAS-domain binding in both NPAS2 and mPER2.

Next, cyanocobalamin was injected i.p. into wild-type mice to investigate its action on the expression of *mPer1* and *mPer2* *in vivo*. Northern blot analysis showed that expression of *mPer1* was suppressed and that of *mPer2* was enhanced by cyanocobalamin in a dose-dependent manner (Fig. 4d). Together, these observations suggest that haem acts as a ligand and that the PAS domain is the ligand-binding motif. The reverse effects on clock gene expression by haem and cobalamin suggest that a second event, perhaps a conformational change mediated through a response to carbon monoxide or a yet unidentified signalling molecule, is required to influence clock gene transcription activity³.

In the fruitfly, the autoregulatory mechanism is modelled on a negative feedback function of dPER¹⁵. In the mouse, CRY proteins are the main negative loop regulators¹⁶, and genetic evidence implicates mPER2 as a positive regulator of the circadian mechanism^{5,17}. Thus, the molecular target and function of mPER2 have been an enigma. Our genetic studies now show that mPER2 is a positive regulator of NPAS2. Reporter assay studies show that mPER2 enhances BMAL1-NPAS2 transcriptional activity. Our studies further show that NPAS2 is a chief regulator of *Alas1* and that haem differentially regulates the expression of *mPer1* and *mPer2* *in vivo*. Together, these findings identify a reciprocal regulation between circadian clock regulation of haem biosynthesis and the haem control of circadian clock-complex-regulated transcription. In this model (Supplementary Fig. 5), expression of *Alas1* is dependent on BMAL1-NPAS2 transcription activity under positive control by mPER2. Through biosynthesis, the abundance of ALAS1 controls that of haem.

Haem is a prosthetic group of both NPAS2 and mPER2. Haem biosynthesis is coordinated with the expression of circadian-controlled genes that encode haemoproteins such as nitric oxide synthase, guanylyl cyclase and other key regulators of biochemical and physiological cascades^{2,18,19}. Increased quantities of haem eventually induce its own degradation through haem-oxygenase enzymes, whose gene expression is induced by free haem²⁰. The degradation products of haem are carbon monoxide, biliverdin and free iron^{19,20}. In the presence of carbon monoxide, the haem-bound BMAL1-NPAS2 transcription complex can no longer bind DNA, stopping gene transcription³. The decline in transcription complex activity is further amplified by a decrease in mPER2. The dissociated transcription complex components are then targeted for degradation. The decline in haem eventually reaches a trough that allows basal quantities of the BMAL1-NPAS2 transcription complex to bind DNA and restart the cycle.

Cell-cycle regulators including c-Myc, Cyclin D, Cyclin A, Cyclin B, Cdc2 and Wee1 are under circadian control *in vivo*^{21,22}. In *Npas2*^{m/m} mice, the c-Myc level is deregulated, as observed in *mPer2*^{m/m} mice²¹ (Supplementary Fig. 6). In line with these observations, haem also affects the expression of c-Myc (Supplementary Fig. 6) and other cell-cycle regulators²³. The proliferation rate of leukaemia cells is affected by analogues of vitamin B12 (ref. 24). Radiation therapy studies have implicated vitamin B12 as a synchronizer of cell division²⁵. Hence, porphyrin molecules are modu-

lators of cell division. Giving vitamin B12 to individuals undergoing pemetrexed chemotherapy reduces drug toxicity²⁶. Pemetrexed is an inhibitor of thymidylate synthase²⁷ whose expression is under circadian control *in vivo*²⁸. Taken together, our studies indicate that the therapeutic use of porphyrin derivatives that target circadian regulators may be beneficial for individuals undergoing chemotherapy and radiation therapy. □

Methods

Animals

Wild-type (C57/B6), *mPer2*^{-/-} (ref. 21), *Npas2*^{m/m} (ref. 10), *mPer1*^{-/-} and *mPer2*^{m/m} (ref. 5) mice were housed in a standard animal maintenance facility under a 12-h/12-h light/dark cycle. We used female mice aged between 2 and 6 months in these studies. These studies were carried out under institutionally approved animal protocol (AN2644).

Northern blot analyses

For all northern blot analyses, 20 µg of total RNA was resolved by gel electrophoresis and transferred onto a nylon membrane. Blots were hybridized with ³²P-labelled cDNA probes for *mPer1*, *mPer2*, *Cry1*, *Bmal1*, *Clock*, *Npas2*, *c-myc* and *Gapdh*, as described^{1,10,11,21}. Hybridization was done at 55 °C in ULTRAhyb (Ambion) hybridization solution; the blots were then washed in 0.2 × SSC, 0.1% SDS at 65 °C and exposed to X-ray films.

Cell culture and haem treatment

NIH 3T3 cells were cultured to a confluent cell density in DMEM medium containing 10% fetal bovine serum (Invitrogen). After 4 d, cells were treated for 2 h in serum-free DMEM containing 30 µM haem (hemin; Sigma). We dissolved haem in solvent as described²⁹. After stimulation, cells were maintained in serum-free DMEM. Cells were collected at 4-h intervals with the first time point at 6 h after haem treatment. RNA was extracted with TRIzol reagent (Invitrogen).

Injection of haem and cyanocobalamin

Haem (30 mg per kg body weight) or solvent was injected i.p. into mice. After injection, the mice were returned to constant darkness and then killed by cervical dislocation 2 h after each of the timed injections. Either cyanocobalamin (Sigma; 10 µg, 150 µg or 2.5 mg) or PBS was injected i.p. into wild-type mice at CT12. After injection, mice were maintained in constant darkness and killed by cervical dislocation at CT16. All injections were done under a 15-W red light. We isolated total RNA from liver tissue and carried out northern blot analysis as described¹.

Locomotor activity

Locomotor activity was monitored as described¹. In brief, for free running activity, mice were entrained to a 12-h/12-h light/dark cycle for 10 d before being released into constant darkness. We injected the mice i.p. with haem or solvent at subjective day (CT6) or subjective night (CT22) for at least two consecutive days under a 15-W red light.

Plasmid constructs

We obtained pCMV-Sport2-mPer2^{m/m} by subcloning a mutant fragment generated by polymerase chain reaction with reverse transcription (RT-PCR) of *mPer2*^{m/m} mouse liver RNA with the oligonucleotides 5'-TACTCGGAGGTAATGCCCTCAACC-3' and 5'-GGCCCTTTGAATGAGGATGTGT-3'. This fragment was used to replace the corresponding fragment of pCMV-Sport2-mPer2 plasmid between the *HindIII* and *BamHI* sites⁵.

To construct the *Alas1* promoter plasmid, the *Alas1* promoter region was isolated from a 3' HPRT mouse genomic library⁵. Using PCR and the oligonucleotides *MluI* 5'-AACCCGACGCGTCCCGCCA-3' (forward) and *XhoI* 5'-CAGTCTCGAGGTTTGCAAA GTGT-3' (reverse), a 3' *Alas1* fragment was generated from the genomic fragment. We cloned the PCR product into the pGL2-basic vector (Promega) at the *MluI* and *XhoI* sites. The 5' region of the *Alas1* promoter was then subcloned into the *MluI* site of the 3' clone to generate a 5,259-base-pair *Alas1* promoter in pGL2-basic. End sequencing of the reporter construct confirmed that the DNA sequence was derived from the *Alas1* promoter region (Celera, clone number MCG19529).

Reporter assays

COS-1 cells (1.3×10^5) were seeded into six-well dishes 24 h before transfection. We transfected cells with 10 ng of the 7.2-kb *mPer1-luc* or the 1.7-kb *mPer2-luc* reporter construct^{8,9}, or with 40 ng of the 5.2-kb *Alas1* reporter construct. Co-transfections were done with the respective expression constructs for *Clock*⁸, *Bmal1*, *Npas2*, *Cry1* (ref. 7), *mPer1* and *mPer2* (ref. 11). The pcDNA 3.1 (Invitrogen) vector was used to bring the total amount of plasmid DNA to 1.6 µg per well. Transfection was done with FuGENE 6 (Roche) and, 24 h later, the cells were collected, lysed and assayed with the Luciferase Assay System (Promega) protocol and a TD-20/201 luminometer (Turner Designs). Total protein concentration was measured with a protein assay (Bio-Rad).

Haem-binding assays

We carried out haem-binding assays as described¹². In brief, COS-1 cells were transfected for 42 h with expression constructs for mPER2, mPER2^{m/m} or NPAS2. Extracts (100 µg of total protein) were incubated in 1 ml of high ionic strength buffer (0.5 M NaCl and 10 mM sodium phosphate, pH 7.5) supplemented with 50 µl of haem-agarose beads (Sigma; containing 335 µM haem) or unconjugated agarose beads at 4 °C for 1 h. Unbound

proteins were removed by washing the beads six times with high ionic strength buffer. Bound protein was eluted with electrophoresis sample buffer containing 1% SDS, 2.5% 2-mercaptoethanol, 0.063 M Tris-HCl (pH 6.8) and 15% glycerol, and resolved by 6% or 8% SDS-PAGE. Western blot analysis was carried out with antisera against mPER2 (ref. 1), NPAS2 (ref. 10) and GAPDH (Sigma). Bound antibodies were detected by Supersignal West Pico Chemiluminescent Substrate (Pierce). Cyanocobalamin (335 μ M) was added to the binding solution for competition studies.

Received 4 March; accepted 3 June 2004; doi:10.1038/nature02724.

- Zheng, B. *et al.* Nonredundant roles of the *mPer1* and *mPer2* genes in the mammalian circadian clock. *Cell* **105**, 683–941 (2001).
- Panda, S. *et al.* Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* **109**, 307–320 (2002).
- Dioum, E. M. *et al.* NPAS2: a gas-responsive transcription factor. *Science* **298**, 2385–2387 (2002).
- Gilles-Gonzalez, M. A. & Gonzalez, G. Signal transduction by haem-containing PAS-domain proteins. *J. Appl. Physiol.* **96**, 774–783 (2004).
- Zheng, B. *et al.* The *mPer2* gene encodes a functional component of the mammalian circadian clock. *Nature* **400**, 169–173 (1999).
- Yamaguchi, S. *et al.* Role of DBP in the circadian oscillatory mechanism. *Mol. Cell. Biol.* **13**, 4773–4781 (2000).
- Rutter, J., Reick, M., Wu, L. C. & McKnight, S. L. Regulation of clock and NPAS2 DNA binding by the redox state of NAD cofactors. *Science* **293**, 510–514 (2001).
- Yamaguchi, S. *et al.* The 5' upstream region of *mPer1* gene contains two promoters and is responsible for circadian oscillation. *Curr. Biol.* **10**, 873–876 (2000).
- Travnickova-Bendova, Z., Cermakian, N., Reppert, S. M. & Sassone-Corsi, P. Bimodal regulation of *mPeriod* promoters by CREB-dependent signaling and CLOCK/BMAL1 activity. *Proc. Natl Acad. Sci. USA* **99**, 7728–7733 (2002).
- Garcia, J. A. *et al.* Impaired cued and contextual memory in NPAS2-deficient mice. *Science* **288**, 2226–2230 (2000).
- Albrecht, U., Sun, Z. S., Eichele, G. & Lee, C. C. A differential response of two putative mammalian circadian regulators, *mPer1* and *mPer2*, to light. *Cell* **91**, 1055–1064 (1997).
- Tsutsui, K. & Mueller, G. C. Affinity chromatography of haem-binding proteins: an improved method for the synthesis of hemin-agarose. *Anal. Biochem.* **121**, 244–250 (1982).
- Hashimoto, S. *et al.* Vitamin B12 enhances the phase-response of circadian melatonin rhythm to a single bright light exposure in humans. *Neurosci. Lett.* **220**, 129–132 (1996).
- Nakamura, T., Uchida, K., Moriguchi, Y., Okamoto, N. & Morita, Y. Transient fluctuation of serum melatonin rhythm is suppressed centrally by vitamin B12. *Chronobiol. Int.* **14**, 549–560 (1997).
- Hardin, P. E., Hall, J. C. & Rosbash, M. Feedback of the *Drosophila* period gene product on circadian cycling of its messenger RNA levels. *Nature* **343**, 536–540 (1990).
- Kume, K. *et al.* mCRY1 and mCRY2 are essential components of the negative limb of the circadian clock feedback loop. *Cell* **98**, 193–205 (1999).
- Shearman, L. P. *et al.* Interacting molecular loops in the mammalian circadian clock. *Science* **288**, 1013–1039 (2000).
- Tuncat, B. *et al.* Circadian variation of nitric oxide synthase activity in mouse tissue. *Chronobiol. Int.* **19**, 393–404 (2002).
- Artinian, L. R., Ding, J. M. & Gillette, M. U. Carbon monoxide and nitric oxide: interacting messengers in muscarinic signaling to the brain's circadian clock. *Exp. Neurol.* **171**, 293–300 (2001).
- Rubio, M. F., Agostino, P. V., Ferreyra, G. A. & Golombek, D. A. Circadian haem oxygenase activity in the hamster suprachiasmatic nuclei. *Neurosci. Lett.* **353**, 9–12 (2003).
- Fu, L., Pelicano, H., Liu, J., Huang, P. & Lee, C. C. The circadian gene *Period2* plays an important role in tumor suppression and DNA damage response *in vivo*. *Cell* **111**, 41–50 (2002).
- Matsuo, T. *et al.* Control mechanism of the circadian clock for timing of cell division *in vivo*. *Science* **302**, 255–259 (2003).
- Zhu, Y., Lee, H. C. & Zhang, L. An examination of haem action in gene expression: haem and haem deficiency affect the expression of diverse genes in erythroid k562 and neuronal PC12 cells. *DNA Cell Biol.* **21**, 333–346 (2002).
- McLean, G. R. *et al.* Cobalamin analogues modulate the growth of leukemia cells *in vitro*. *Cancer Res.* **57**, 4015–4022 (1997).
- Levitman, M. K. *et al.* Antitumor effect of combined treatment with ionizing radiation and vitamin B12–C complex. *Radiat. Biol. Radioecol.* **42**, 511–514 (2002).
- Scagliotti, G. V. *et al.* Phase II study of pemetrexed with and without folic acid and vitamin B12 as front-line therapy in malignant pleural mesothelioma. *J. Clin. Oncol.* **21**, 1556–1561 (2003).
- Norman, P. Pemetrexed disodium (Eli Lilly). *Curr. Opin. Invest. Drugs* **2**, 1611–1622 (2001).
- Lincoln, D. W. II, Hrushesky, W. J. & Wood, P. A. Circadian organization of thymidylate synthase activity in normal tissues: a possible basis for 5-fluorouracil chronotherapeutic advantage. *Int. J. Cancer* **88**, 479–485 (2000).
- Freedman, M. L., Geraghty, M. & Rosman, J. Hemin control of globin synthesis. Isolation of a hemin-reversible translational repressor from human mature erythrocytes. *J. Biol. Chem.* **249**, 7290–7294 (1974).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. L. McKnight for the *Npas2*^{tm/m} mice and NPAS2 antibody; M. Reick for technical comments; H. Okamura for the *mPer1-luc* and *Clock* expression constructs; P. Sassone-Corsi for the *mPer2-luc* expression plasmid; Z. Sun for suggesting the use of haem-agarose; and P. Hastings, R. Kellemis and J. Lever for comments. This work was supported, in part, by a grant from the NIH (to C.C.L.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.C.L. (cheng.c.lee@uth.tmc.edu).

Role of transposable elements in heterochromatin and epigenetic control

Zachary Lippman^{1*}, Anne-Valérie Gendrel^{2*}, Michael Black^{3†}, Matthew W. Vaughn¹, Neilay Dedhia¹, W. Richard McCombie¹, Kimberly Lavine¹, Vivek Mittal¹, Bruce May¹, Kristin D. Kasschau⁴, James C. Carrington⁴, Rebecca W. Doerge³, Vincent Colot² & Rob Martienssen¹

¹Watson School of Biological Sciences and Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

²Unité de Recherche en Génétique Végétale (URGV), INRA/CNRS/UEVE, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France

³Department of Statistics, Purdue University, West Lafayette, Indiana 47907, USA

⁴Center for Gene Research and Biotechnology, Oregon State University, Corvallis, Oregon 97330, USA

* These authors contributed equally to this work

† Present address: Department of Statistics, The University of Auckland, Private Bag 92019, Auckland, New Zealand

Heterochromatin has been defined as deeply staining chromosomal material that remains condensed in interphase, whereas euchromatin undergoes de-condensation¹. Heterochromatin is found near centromeres and telomeres, but interstitial sites of heterochromatin (knobs) are common in plant genomes and were first described in maize². These regions are repetitive and late-replicating³. In *Drosophila*, heterochromatin influences gene expression, a heterochromatin phenomenon called position effect variegation⁴. Similarities between position effect variegation in *Drosophila* and gene silencing in maize mediated by “controlling elements” (that is, transposable elements) led in part to the proposal that heterochromatin is composed of transposable elements, and that such elements scattered throughout the genome might regulate development². Using microarray analysis, we show that heterochromatin in *Arabidopsis* is determined by transposable elements and related tandem repeats, under the control of the chromatin remodelling ATPase DDM1 (Decrease in DNA Methylation 1). Small interfering RNAs (siRNAs) correspond to these sequences, suggesting a role in guiding DDM1. We also show that transposable elements can regulate genes epigenetically, but only when inserted within or very close to them. This probably accounts for the regulation by DDM1 and the DNA methyltransferase MET1 of the euchromatic, imprinted gene *FWA*, as its promoter is provided by transposable-element-derived tandem repeats that are associated with siRNAs.

Arabidopsis heterochromatin is organized into chromocentres in interphase nuclei⁵. It is localized mostly to the pericentromeric and nucleolar organizing regions (NOR), and two domains of interstitial heterochromatin (knobs) on chromosomes 4 (*hk4S*) and 5 have been completely sequenced^{5–7}. These regions are rich in transposable elements, but in each case the most conspicuous feature is a tandem array of ~2-kilobase (kb) satellite repeats. The arrays are unrelated and not found elsewhere, although they have limited homology with CACTA and *Mutator-like* (MULE) DNA transposons, respectively. *hk4S* is part of a duplicated segment found on the long arm of chromosome 4 (ref. 7) and encompasses 8 of the 33 genes found in conserved order and orientation (Fig. 1). However, although the long arm segment is virtually free of transposable elements, the heterochromatic segment is interrupted by at least 34 retrotransposons and 40 DNA class transposons⁸, often inserted into one another. All but one of these insertions arose after the duplication event, indicating that heterochromatin was derived

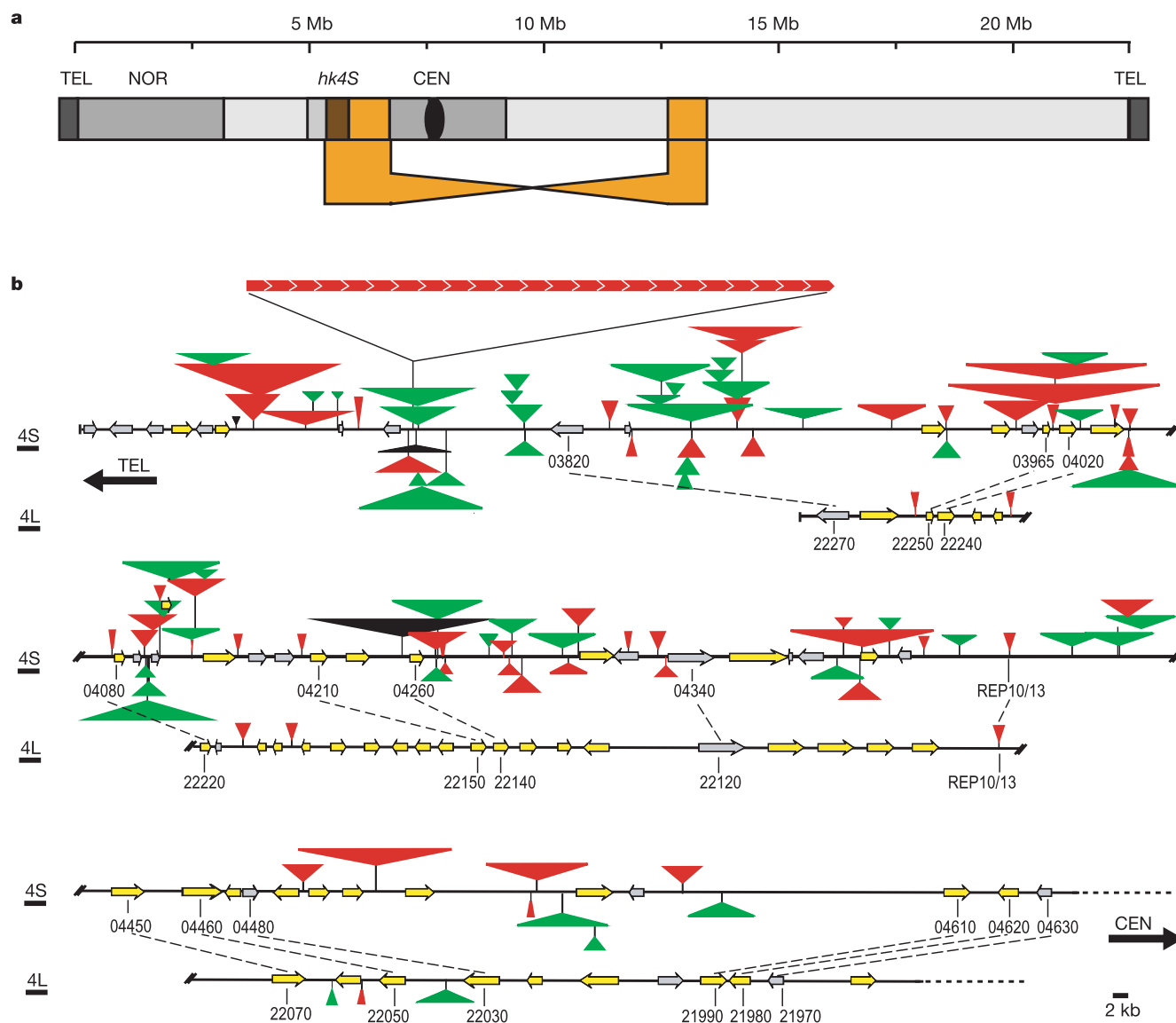


Figure 1 The heterochromatic knob (*hk4S*) on chromosome 4. **a**, The chromosomal position of the segmental duplication between 4S and 4L (orange) is indicated relative to heterochromatin (grey), and includes part of *hk4S* (brown). The segments are inverted around the centromere (CEN). NOR, nucleolar organizing region; TEL, telomere. **b**, Genome organization of the segmentally duplicated regions. Transposable elements are shown as nested insertions between known genes (yellow arrows) and hypothetical genes (grey arrows), and arose after the region was duplicated. DNA transposons and

retrotransposons are indicated by red and green triangles, respectively. Black triangles represent complex transposable element insertions that we were unable to resolve. The orientation of each region relative to the telomere and centromere is shown. Homologous genes between the duplicated regions are in collinear order and are indicated by locus designators starting with At4g03820 (03820); they are joined by dashed lines.

Figure 2 Expression and chromatin profiling of wild type and *ddm1* using genomic tiling microarrays. **a**, The entire 1.5-Mb region surrounding *hk4S*. The region between 1,200–2,670 kb from the NOR at the tip of chromosome 4 is displayed along with coordinates (in kb). *hk4S* lies between 1,600–2,330 kb (black bar). Relative transcription levels (mRNA) are indicated for wild-type (green) and *ddm1* mutant seedlings (red). Annotation includes four tracks (from top to bottom): 1, ORFs; 2, repeats; 3, gene trap insertions; and 4, small RNA. ORFs (TIGR v.2.0) are annotated as known genes (yellow), hypothetical genes (grey), retrotransposons (green) and DNA transposons (red) (track 1). Repeats are from RepBase[®] with long terminal repeats (LTRs) indicated by pink boxes (see Methods) (track 2). Gene trap insertions are indicated by blue lines (track 3); whereas small RNA matches are indicated by black lines (track 4). DNA methylation (5mC), histone H3 methylation of lysine 9 (mK9), and histone H3 methylation of lysine 4 (mK4) that was significantly above or below the average level found in euchromatic features are highlighted in brown (5mC), blue (mK9) and green (mK4), respectively, for

wild type (WT) and *ddm1* (Methods). Those features with euchromatic levels of each modification are coloured grey. Blank tiles reflect failure to amplify and print DNA. Examples of gene islands are highlighted by boxes at 1,770, 1,865, 1,930, 2,020 and 2,115 kb. These regions are expressed equally in *ddm1* and wild type, and have reduced 5mC and mK9. The BAC T27D20, which decorates a euchromatic loop in interphase cells⁵, spans the region from 2,001 to 2,081 kb. **b**, Epigenetic activation of heterochromatin. Expression profiling of *ddm1* mutants (top) is compared to *ddm1*/+ plants (bottom) after backcrosses, compared with wild-type plants in each case. Similar profiles indicate that epigenetically activated heterochromatin is inherited. The *ATGPI* element at 1,620 kb is silenced in this backcross. **c**, Genes are insulated in gene islands. A close-up of the gene island at 1,930 kb and specific PCR amplification of reverse-transcribed cDNA (RT-PCR) for select genes and transposable elements (+). Mock RT-PCR was performed without reverse transcriptase (–). LINE, long interspersed nuclear element.

from euchromatin by the insertion of transposable elements (Fig. 1b).

Typically, euchromatin is associated with acetylated histones and with histone H3 di-methylation on lysine 4 (H3mK4), whereas heterochromatin is associated with histone H3 di-methylation on lysine 9 (H3mK9)^{9,10}. In many eukaryotes, heterochromatic DNA is also heavily methylated¹¹. Mutations in the *Arabidopsis* chromatin remodelling gene *DDM1* affect both DNA and histone H3 methylation, perhaps because chromatin-modifying enzymes act in multi-protein complexes¹². *DDM1* is conserved between yeast, animals and plants, and mutants have comparable phenotypes in the mouse¹³.

To determine the sequences responsible for heterochromatic modifications in *Arabidopsis*, we profiled histone and DNA modifications in *hk4S*, as well as expression, using DNA microarrays. Silent heterochromatic elements are poorly represented on commercial microarrays; therefore, we took a tiling approach by amplifying sequential 1-kb segments over a 1.5-megabase (Mb) region centred on *hk4S*. In addition, 36 unlinked 10-kb regions were included, either as controls or because of their potential regulation by *DDM1*. Microarrays were printed on glass slides and hybridized with labelled complementary DNA, or with labelled genomic DNA that had been enriched for sequences recovered by chromatin immunoprecipitation (ChIP) with antibodies raised against H3mK4 or H3mK9 (Supplementary Information), or depleted of methylated sequences (Supplementary Fig. S1). Control hybridizations with total genomic DNA allowed DNA methylation and histone H3 methylation to be quantified for each feature on the array.

A linear model approach was used to estimate technical and biological effects, allowing statistical detection of features significantly enriched for cytosine methylation and histone modifications, and to detect differential gene expression (see Methods). Individual results were validated by polymerase chain reaction (PCR) with specific primer pairs to account for cross-hybridization with closely related sequences elsewhere in the genome (Supplementary Fig. S4). The microarray data were displayed graphically in the context of genomic annotation (Fig. 2) using a customized version of Generic Genome Browser (<http://chromatin.cshl.org/ddm1/>).

As expected, sequences from *hk4S* had high levels of H3mK9 and DNA methylation, but were significantly depleted of H3mK4 relative to the surrounding euchromatin (Fig. 2a). Notably, individual features associated with high levels of H3mK9 were almost always methylated, indicating a genome-wide relationship between these heterochromatic marks. Moreover, H3mK9 and DNA methylation were not distributed uniformly, but coincided with transposable elements and related repeats. DNA methylation was markedly reduced in *ddm1* mutants, whereas H3mK9 was lost from heterochromatin. Erasure of heterochromatic marks was accompanied by a rather uniform replacement with H3mK4 to levels typical of euchromatin. H3mK9 and H3mK4 are both retained in *ddm1* mutants (ref. 14), suggesting that they are re-distributed rather than lost, consistent with a role for *DDM1* in heterochromatic histone variant exchange¹⁰.

Transposable elements were associated with heterochromatic modification both within and outside *hk4S*. For example, *VANDAL2* (At4g03300) and *ATENSPM4* (At4g03310) are inserted in euchromatin and are associated with H3mK9 and DNA methylation (at 1,450 and 1,459 kb, respectively, in Fig. 2a). Thus, these elements constitute 'cryptic' heterochromatin, defined biochemically but not cytologically. In wild type, transposable elements were silent or expressed at low levels, but in *ddm1* more than half were strongly expressed (Fig. 3). Often, transposable elements that were disrupted by other transposable elements permanently lost their activity (Supplementary Fig. S3). Expressed transposable elements were inserted more recently and included both low- and high-copy elements such as CACTA transposons and Athila gypsy-like retrotransposons,

respectively. Thus copy number did not account for high levels of expression in *ddm1*. In wild type, CACTA and gypsy-like transposable elements were the most uniformly methylated and associated with H3mK9 (Supplementary Fig. S3 and Table S1).

ddm1-2 is strictly recessive, but, when crossed to wild-type, hypomethylated centromeric and rDNA sequences were not re-methylated¹⁵. To investigate whether activated heterochromatin is inherited in subsequent generations, *ddm1* was out-crossed to wild type, and messenger RNA was extracted from the resulting *ddm1*/+ plants. Expression profiles were similar to *ddm1*, indicating that transposable elements activated in *ddm1* plants retained their activity in *ddm1*/+ plants even though *DDM1* function was restored (Fig. 2b). The gypsy class element *ATGP1* was exceptional in that it was partially silenced in the backcross (at 1,620 kb in Fig. 2b). Interestingly, DNA methylation, H3mK9 and siRNA were retained at *ATGP1* in *ddm1* plants (ref. 12).

One hundred and six known genes are present in the 1.5-Mb region centred on *hk4S*; 16 of these lie in the knob, along with 15 hypothetical open reading frames (ORFs). Most genes were expressed, together with contiguous unannotated features, representing 5' and 3' untranslated regions. Unlike transposable elements, genes were expressed at similar levels in wild-type and *ddm1* plants, regardless of location. Genes in the knob are found in gene islands (Fig. 2a, c), which may correspond to interphase 'loops'

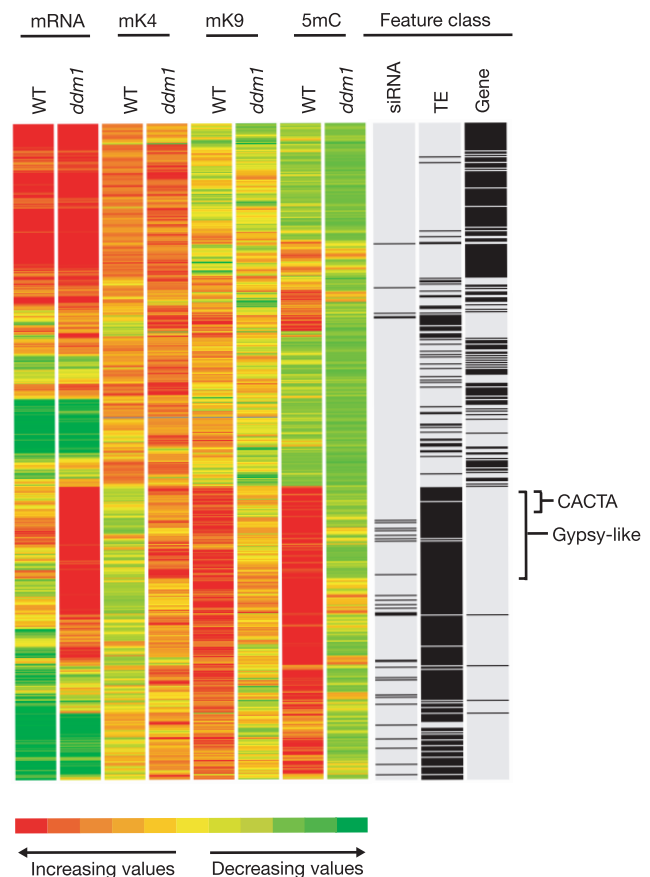


Figure 3 Cluster analysis. Tiles corresponding to ORFs were clustered according to the levels of expression (mRNA), H3mK4 (mK4), H3mK9 (mK9) and DNA methylation (5mC) in wild-type and *ddm1* plants using GeneSpring 6.0, and subsequently scored as matching known genes (Gene), transposable elements (TE) and siRNA (siRNA). Genes and transposable elements were sharply distinguished from each other. Genes were expressed in wild type and associated with gene traps and with H3mK4, whereas transposable elements were silent, and associated with siRNA, H3mK9 and DNA methylation. In *ddm1*, transposable elements adopted gene-like chromatin properties and the majority were expressed.

of euchromatin that emanate from the chromocentre⁵. Most genes on the array were unmethylated, although low levels of methylation were detected in some genes, regardless of their expression (Supplementary Table S1). For example, two closely related callose synthase genes (at 1,572 and 2,536 kb in Fig. 2) were methylated at their 3' ends (Supplementary Fig. S2). Unlike transposable elements, methylated genes were not associated with significant levels of H3mK9 in wild type, and their expression (Fig. 2a) and DNA methylation was usually unchanged in *ddm1* (Supplementary Fig. S2). We conclude that genes are insulated from DDM1-mediated heterochromatic silencing.

Gene and enhancer traps are sensitive indicators of chromatin

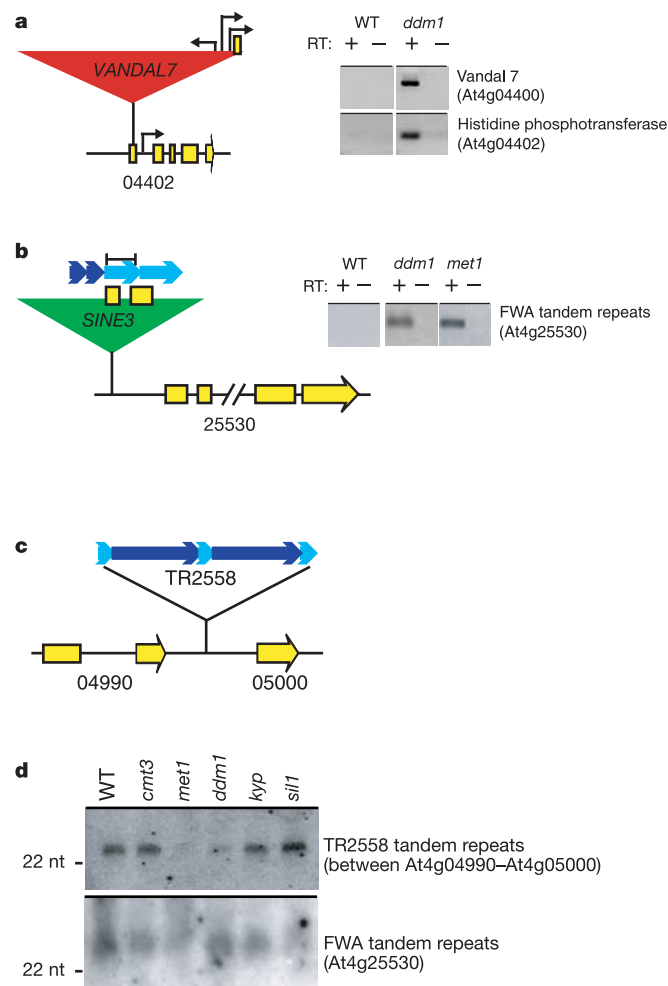


Figure 4 DDM1-dependent gene regulation. **a**, The histidine phosphotransferase gene At4g04402. This gene is interrupted by a *MuDR*-like *VANDAL* transposon and is silenced in wild type. RT-PCR using specific primers indicates that both the gene and the transposon are activated in *ddm1*. Rapid amplification of cloned cDNA ends (5' RACE) was used to map transcription start sites (arrows). The *VANDAL* element provides two of the three sites detected for AT4g04402. **b**, *FWA* and its associated repeats. The promoter and upstream region of the late-flowering gene *FWA* are provided by a SINE retrotransposon and associated tandem repeats (a detailed analysis is shown in Supplementary Fig. S5). These repeats are silent in wild type, but transcripts accumulate in *ddm1* and *met1*. Tandem repeat transcripts are multimeric (data not shown). The primer set used for RT-PCR is shown (brackets). **c**, Tandem repeats corresponding to siRNA. The arrangement of intergenic tandem repeats at coordinate 2,558 kb (*TR2558*) is shown. **d**, siRNA corresponding to tandem repeats were reduced in *ddm1* and/or *met1*. For comparison, chromatin mutants with little effect on siRNA accumulation are shown: the non-CG DNA methyltransferase *cmt3* (*chromomethylase3*), the histone lysine 9 methyltransferase *kyp* (*kryptonite*), and the histone deacetylase *sil1* (*silencing locus1*). A 22-oligonucleotide (nt) size marker is indicated.

structure⁶. Among the 3,000 insertions mapped to the short arm of chromosome 4, there is a pronounced bias for recovery of insertions in gene islands (<http://genetrap.cshl.org>). For example, there are six gene trap insertions in At4g04020 (at 1,931 kb in Fig. 2a) but none in the surrounding 100 kb, indicating that genes within the knob are more permissive than transposable elements for marker expression.

Although genes and transposable elements are sharply distinguished by DDM1, they resemble each other in most respects, such as G+C content and coding capacity. Thus, the mechanism by which they are differentiated is probably sequence specific. One possibility is that transposable elements are recognized by RNA interference (RNAi). RNAi silences transposable elements in many eukaryotes, generating 21–24-nucleotide siRNA derived from double-stranded (ds)RNA by the RNase-helicase Dicer^{16,17}. siRNAs from *Arabidopsis* are currently being cloned and sequenced (<http://cgrb.orst.edu/smallRNA/db/>), and so far 93 features from the 1.5-Mb region match one or more siRNA (Figs 2 and 3; <http://chromatin.cshl.org/ddm1/>).

To determine the role of DDM1 and siRNA in targeting transposable elements, microarray data were subjected to unsupervised cluster analysis and then compared to annotation (Supplementary Methods). Features annotated as transposable elements and siRNA clustered together, whereas genes were more similar to each other (Fig. 3). Although cluster analysis is a useful exploratory tool, confidence levels cannot be easily assigned to any given association. Therefore, significantly enriched or depleted features from the linear model analysis were also summarized (Fig. 2a; see also Supplementary Table S1). Transposable elements and siRNAs were associated with H3mK9 and DNA methylation, and transposable elements that corresponded to siRNA were typically expressed in *ddm1* mutants (Fig. 3; see also Supplementary Table S1). This observation suggests that transcription of transposable elements targets their heterochromatic modification, consistent with a mechanism whereby siRNA guides DNA and histone methylation¹⁸. dsRNA can arise from transposable elements by read-through transcription, and transcripts were detected from both strands of many transposable elements in *ddm1*, probably due to integration into other transposable elements (Supplementary Fig. S3). This mechanism can account for much of the siRNA specific to transposable elements^{16,17}.

Tandem repeats are conspicuous features of heterochromatin, and theoretically they can amplify siRNA through re-iterative RNA replication and cleavage¹⁹. The knob sequence includes 22.5 copies of a 1,850-base-pair (bp) tandem repeat, *AtENSAT1* (1,700–1,750 kb in Fig. 2a), of which the first 300 bp correspond to the terminal inverted repeats of the CACTA transposon *ATENSPM4* (refs 6, 8). In addition, there are several euchromatic tandem repeats represented on the array including *TR2558* (at 2,558 kb in Fig. 2a) (Fig. 4c). Both *ATENSAT1* and *TR2558* accumulate a large number of siRNAs. The repeats are methylated and associated with H3mK9. These modifications are under the control of DDM1, and siRNAs from *TR2558* are severely reduced in both *ddm1* and *met1* mutants (Fig. 4d). Thus siRNAs are derived from both transposable elements and tandem repeats, and are regulated by DDM1.

The high resolution provided by the microarray (~1 kb) indicated that DDM1 sharply distinguishes genes and transposable elements (Fig. 2); however, a handful of genes found in the knob were silenced by DDM1. At4g04402 is related to histidine phosphotransferase *AtHP2* genes found on chromosomes 3 and 5. Although these paralogues were strongly expressed and associated with H3mK4 (Supplementary Fig. S4), At4g04402 was silent, heavily methylated and associated with H3mK9 (at 2,163 kb in Fig. 2a). A 13-kb *VANDAL7* transposon is inserted 134 bp within the silent gene, but in *ddm1* mutants both the transposon and the gene were transcriptionally activated (Fig. 4a). Furthermore, this activation was inherited in *ddm1/+* plants (data not shown).

VANDAL elements belong to the *MuDR* superfamily, which have

been shown to bring genes under epigenetic control in maize²⁰. As in maize, At4g04402 transcripts arose from an outward-reading promoter provided by the transposable element (Fig. 4a). In contrast, a similar VANDAL7 element integrated less than 500 bp upstream of At4g04320 (at 2,112 kb in Fig. 2a) did not bring the gene under its control. Three other genes in the knob (At4g03950, At4g04080 and At4g04110) were silenced by DDM1, and had repetitive insertions relative to euchromatic paralogues (data not shown). However, activation was not observed in every biological replicate, and was not explored further. We conclude that epigenetic gene silencing mediated by transposable elements requires insertion within or very near the gene.

Heritable changes in the expression of transposable-element-regulated genes provide a potential mechanism for the 'syndrome' of sporadic phenotypes that arise in inbred *ddm1* mutants¹⁵. One such phenotype is a delay in flowering owing to inappropriate expression of *FWA*, which encodes a homeodomain protein normally expressed in the seed²¹. *FWA* expression is imprinted in the endosperm under the control of *MET1* (ref. 22). Although it is not located in the knob, we included *FWA* on the array because of its regulation by DDM1 (ref. 21).

The *FWA* promoter and 5' untranslated region comprise two sets of tandem repeats of 38 bp and 198 bp, respectively, which are methylated under DDM1 control²¹. Because of the involvement of transposable elements in DDM1-mediated gene silencing, we inspected this region and discovered that the repeats are encompassed within a 422-bp *SINE3* retrotransposon, distantly related to *Arabidopsis* *AtSN2* (Supplementary Fig. S5). Short interspersed nuclear element (SINE) insertion and additional rearrangements involving a 10-bp target site duplication result in the retroelement providing the first two exons of the gene (Supplementary Fig. S5), including the start site of *FWA* transcription²¹. Microarray analysis revealed that this retrotransposon is associated with H3mK9 in wild type but not in *ddm1* (<http://chromatin.cshl.org/ddm1/>). These data indicate that *FWA* is under transposable element control, and implicate transposable elements in imprinting. To explore this mechanism further, we searched for siRNAs corresponding to the 198-bp repeat. These were detected in wild type and *ddm1*, but were reduced in *met1* (Fig. 4d). Low levels of transcripts corresponding to both copies of the repeat could be readily detected in *ddm1* and *met1*, and are probably the source of these siRNAs (Fig. 4b).

We have found that heterochromatin, whether cytologically visible or not, is defined by transposable elements, consistent with the view expressed by B. McClintock more than 50 years ago². Genes are insulated from heterochromatin within local euchromatic environments⁵. A similar organization is found within a rice centromere²³, and may also apply to the bulk of the maize genome, given that gene islands are unmethylated and separated by a sea of methylated transposable elements²⁴.

In fission yeast and in *Drosophila*, which mostly lack DNA methylation, RNA interference and H3mK9 are required for position effect variegation associated with heterochromatic repeats^{10,18,25}. Position effect variegation silences nearby genes, but the silencing is relatively unstable, resulting in variegation. In *Arabidopsis*, both DNA methylation and H3mK9 are maintained by DDM1, and heterochromatic silencing is heritable. DDM1 distinguishes genes from transposable elements and repeats, and siRNAs probably contribute to this distinction. Some siRNAs corresponding to transposable elements are lost in *ddm1* mutants, suggesting that siRNAs are engaged in a stabilizing interaction with DDM1 (ref. 12). In fission yeast, the H3K9 methyltransferase *clr4* + is similarly required for siRNA accumulation²⁶, possibly mediated by the RITS complex²⁷. However, the *Arabidopsis* RNAi genes *AGO1* and *AGO4* have only limited roles in transposable element silencing^{12,28}. One explanation is that siRNAs are only needed to target heterochromatic modifications, which once established, are maintained by DDM1. This would also explain why

silencing cannot be re-established on out-crossing of *ddm1* once H3mK9, DNA methylation and siRNA are lost¹².

We show that transposable elements silence genes epigenetically, but only when they integrate within or very near them, resembling suppressible insertions in plants, animals and fungi¹¹. In the mouse, for example, the insertion of an IAP retrotransposon immediately upstream of the agouti gene provides new transcripts and leads to mutant phenotypes only when unmethylated.

Late-flowering epimutants of the euchromatic gene *FWA* arise in *ddm1* and even more frequently in *met1* (refs 15, 21). Furthermore, *FWA* expression is imprinted in the endosperm under the control of *MET1* (ref. 22). The promoter and first two exons of *FWA* are provided by a SINE, which brings the gene under the control of DDM1. The SINE contains tandem repeats that are methylated and correspond to siRNA. As with heterochromatic tandem repeats, these siRNAs are reduced in *met1*, and longer transcripts accumulate in both *met1* and *ddm1*. Loss of tandem repeat siRNA in the female germ line might account for loss of heterochromatic modifications and specific expression from the maternal allele in the developing seed. Notably, differentially methylated regions of imprinted genes in mammals are also transcribed and contain tandem repeats, some of which correspond to microRNA²⁹. Thus, in repeat-rich genomes, transposable elements and related repeats are likely to have major regulatory roles in development and disease. *Note added in proof:* While under review, additional reports^{30,31} have indicated that RNAi is required to establish *FWA* transgene silencing in *Arabidopsis*, in agreement with our results. □

Methods

Plant material

All wild-type and homozygous *ddm1-2* *Arabidopsis* plants used in this study were of the Columbia ecotype. Individual *ddm1-2* mutant lines were self-pollinated for three generations and then pooled for analysis¹⁴. Plants were grown *in vitro* on 0.5 × Murashige and Skoog (MS) media, with sucrose (0.7%). Nine-day-old seedlings were then collected for chromatin immunoprecipitation, DNA and RNA extraction. Endogenous siRNAs were identified from cDNA libraries prepared using small RNA isolated from inflorescence tissues. All siRNAs can be viewed at <http://cgrb.orst.edu/smallRNA/db/>. Small RNA detection in chromatin mutants was carried out as previously described¹² using soil-grown plants in the Landsberg *erecta* background. The blot shown in Fig. 4d was repeated twice using microRNA miR171 as a loading control (not shown). It was unaffected in all genotypes¹².

Microarrays

Bacterial artificial chromosome (BAC) templates covering the chromosome 4 heterochromatic knob and flanking sequences were used to generate the fragments printed on microarrays by PCR amplification using primers selected at 1-kb intervals. BACs come from the 'tiling path' used for sequencing of the genome of *Arabidopsis thaliana* cv. Columbia (<http://www.mips.biochem.mpg.de/proj/thal/db/>). The 20 BACs used were: T10P11, T5J8, T4I9, F4C21, F9H3, T5L23, T5H22, T7M24, T25H8, T24M8, T24H24, T27D20, T19B17, T26N6, F4H6, T19J18, T4B21, T1J1, T32N4 and C17L7. All PCR products were quantified on agarose gels, before printing in duplicate or triplicate onto glass slides. Polyadenylated RNA from wild-type and *ddm1* seedlings was labelled indirectly with Cy3 and Cy5 and hybridized to pairs of slides in dye-swap experiments. Two dye swaps were performed using the same RNA (that is, four technical replicates), and one dye swap was carried out using seedlings grown under similar conditions (two biological replicates). In addition to labelled cDNA, the arrays were also hybridized with total genomic DNA extracted from seedlings and the same genomic DNA depleted for methylated sequences. DNA was sheared to a constant size using nebulization and then divided into two equal samples, one of which was digested with the methylation-dependent restriction enzyme MspI, which cuts the sequence A/G 5mC. Depleted and undigested DNA samples were then size-fractionated on agarose gels to recover fragments greater than 1 kb, closely matching the resolution provided by the tiling array. The intensity ratios gave a quantitative indication of cytosine methylation. We performed chromatin immunoprecipitation with antibodies raised against H3 dimethyl K4 or H3 dimethyl K9 (Upstate Biotechnology). Whole seedlings were treated with formaldehyde to fix histone-DNA complexes, chromatin was extracted, sonicated and immunoprecipitated before amplification, labelling and hybridization to the array.

A linear model approach³² was used to estimate experimental effects, determine the level of variability, and detect features undergoing statistically significant changes in fluorescence intensity between the conditions investigated by each experiment. The linear models procedure partitions the sources of variation such that global and feature-specific array and dye effects are removed, creating a corrected signal for each feature. This signal was then used to assess changes in fluorescence intensity for each feature solely due to differences between the conditions of interest. In the case of DNA methylation the average ratio of hybridization intensities for selected DNA and total DNA was calculated for each feature in the dye-swap experiments. In both wild-type and *ddm1* samples, unmethylated

features gave a ratio of close to 1.0, whereas methylated features gave a ratio greater than 1.0. Similarly, for chromatin immunoprecipitation, the ratio of selected and total DNA was calculated for each feature. In this case the ratio found in euchromatin was arbitrarily set to 1.0 in both mutant and wild type, as recovery of immunoselected DNA is always much less than 100%. This is the equivalent of using euchromatic genes such as actin as a control. Statistical tests were used to detect features undergoing significant changes, based on control of both the family-wise error rate (FWER) and the false discovery rate (FDR)³².

This statistical analysis allows widely differing profiles to be compared such that the resulting data were highly reproducible, with less than 1% of the features displaying different expression ratios under biological replication (not shown). Because of the repetitive nature of the sequences involved, cross-hybridization between repeats and gene family members was likely to account for some of the signal observed. However, PCR validation with specific primer pairs indicated that, except in rare instances, hybridization provided an accurate measure of expression as well as DNA and histone methylation patterns across the 1.5-Mb region represented on the array (Supplementary Methods).

Received 2 January; accepted 7 May 2004; doi:10.1038/nature02651.

- Heitz, E. Das heterochromatin der Moose. *Jahrb. Wiss. Botanik* **69**, 762–818 (1928).
- McClintock, B. Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* **16**, 13–47 (1951).
- Hennig, W. Heterochromatin. *Chromosoma* **108**, 1–9 (1999).
- Schotta, G., Ebert, A., Dorn, R. & Reuter, G. Position-effect variegation and the genetic dissection of chromatin regulation in *Drosophila*. *Semin. Cell Dev. Biol.* **14**, 67–75 (2003).
- Franz, P., De Jong, J. H., Lysak, M., Castiglione, M. R. & Schubert, I. Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. *Proc. Natl Acad. Sci. USA* **99**, 14584–14589 (2002).
- CSHL/WUGSC/PEB, *Arabidopsis* Sequencing Consortium. The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100**, 377–386 (2000).
- Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
- Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
- Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
- Vermaak, D., Ahmad, K. & Henikoff, S. Maintenance of chromatin states: an open-and-shut case. *Curr. Opin. Cell Biol.* **15**, 266–274 (2003).
- Martienssen, R. A. & Colot, V. DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293**, 1070–1074 (2001).
- Lippman, Z., May, B., Yordan, C., Singer, T. & Martienssen, R. Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. *PLoS Biol.* **1**, E67 (2003).
- Verbsky, M. L. & Richards, E. J. Chromatin remodeling in plants. *Curr. Opin. Plant Biol.* **4**, 494–500 (2001).
- Gendrel, A. V., Lippman, Z., Yordan, C., Colot, V. & Martienssen, R. A. Dependence of heterochromatic histone H3 methylation patterns on the *Arabidopsis* gene DDM1. *Science* **297**, 1871–1873 (2002).
- Kakutani, T. Epi-alleles in plants: inheritance of epigenetic information over generations. *Plant Cell Physiol.* **43**, 1106–1111 (2002).
- Sijen, T. & Plasterk, R. H. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* **426**, 310–314 (2003).
- Aravin, A. *et al.* Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.* **11**, 1017–1027 (2001).
- Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
- Martienssen, R. A. Maintenance of heterochromatin by RNA interference of tandem repeats. *Nature Genet.* **35**, 213–214 (2003).
- Barkan, A. & Martienssen, R. A. Inactivation of maize transposon Mu suppresses a mutant phenotype by activating an outward-reading promoter near the end of Mu1. *Proc. Natl Acad. Sci. USA* **88**, 3502–3506 (1991).
- Soppe, W. J. *et al.* The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol. Cell* **6**, 791–802 (2000).
- Kinoshita, T. *et al.* One-way control of FWA imprinting in *Arabidopsis* endosperm by DNA methylation. *Science* **303**, 521–523 (2004).
- Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nature Genet.* **36**, 138–145 (2004).
- Rabinowicz, P. D. *et al.* Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res.* **13**, 2658–2664 (2003).
- Pal-Bhadra, M. *et al.* Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science* **303**, 669–672 (2004).
- Schramke, V. & Allshire, R. Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science* **301**, 1069–1074 (2003).
- Verdel, A. *et al.* RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303**, 672–676 (2004).
- Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299**, 716–719 (2003).
- Seitz, H. *et al.* Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. *Nature Genet.* **34**, 261–262 (2003).
- Chan, S. W. *et al.* RNA silencing genes control de novo DNA methylation. *Science* **303**, 1336 (2004).
- Xie, Z. *et al.* Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2**, E104 (2004).
- Craig, B. A., Black, M. A. & Doerge, R. W. Gene expression data: the technology and statistical analysis. *J. Agric. Biol. Environ. Stat.* **8**, 1–28 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. Richards and our colleagues T. Osborn, L. Comai, J. Chen and J. Birchler for their comments and advice. We also thank P. Rabinowicz for advice on ChIP microarray experiments. V.C. thanks M. Caboche for laboratory space and continuous support. Z.L. is an Arnold and Mabel Beckman graduate fellow in the Watson School of Biological Sciences. A.V.G. is supported by a graduate studentship from the French Ministry of Research. M.V. is a National Science Foundation Bioinformatics postdoctoral fellow. This work was supported by a grant from the NSF Plant Genome Program (to R.W.D. and R.M.), as well as grants from Genopole and the CNRS (to V.C.), grants from NSF and NIH to J. C., and NIH to R.M.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to R.M. (martiens@cshl.edu) or V.C. (colot@evry.inra.fr).

Periodic cycles of RNA unwinding and pausing by hepatitis C virus NS3 helicase

Victor Serebrov¹ & Anna Marie Pyle^{1,2}

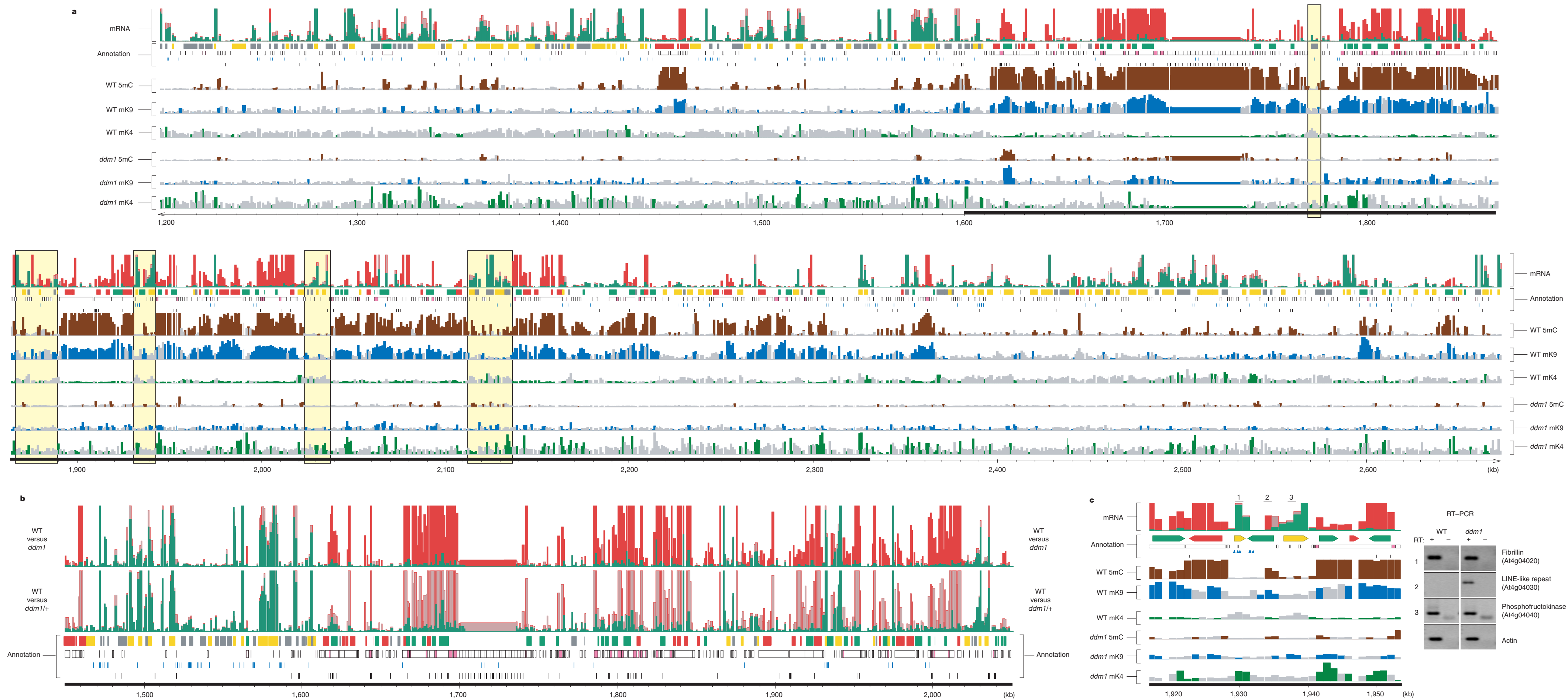
¹Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

²Howard Hughes Medical Institute, New Haven, Connecticut 06520, USA

The NS3 helicase is essential for cytoplasmic RNA replication by the hepatitis C virus^{1–4}, and it is a representative member of helicase superfamily 2 (SF2). NS3 is an important model system for understanding unwinding activities of DExH/D proteins^{5–7}, and it has been the subject of extensive structural and mutational analyses^{8–11}. Despite intense interest in NS3, the molecular and kinetic mechanisms for RNA unwinding by this helicase have remained obscure. We have developed a combinatorial, time-resolved approach for monitoring the microscopic behaviour of a helicase at each nucleotide of a duplex substrate. By applying this analysis to NS3, we have independently established the ‘physical’ and ‘kinetic’ step size for unwinding of RNA (18 base pairs, in each case), which we relate to the stoichiometry of the functional, translocating species. Having obtained microscopic unwinding rate constants at each position along the duplex, we demonstrate that NS3 unwinds RNA through a highly coordinated cycle of fast ripping and local pausing that occurs with regular spacing along the duplex substrate, much like the stepping behaviour of cytoskeletal motor proteins¹².

Replication by the hepatitis C virus (HCV) is performed by a protein complex that includes the multifunctional NS3 helicase, the NS5B polymerase and other nonstructural proteins^{1–4}. Although there have been numerous studies on the protease, ATPase and helicase activities of the protein^{13–16}, a quantitative analysis of its RNA unwinding mechanism has never been conducted. Helicase activity is traditionally monitored on short, ‘tailed’ substrates that consist of a duplex flanked by single stranded nucleic acid. Information about helicase mechanism (such as kinetic parameters and processivity) can be obtained by monitoring the extent and efficiency of unwinding for a series of duplexes that sequentially increase in length. Although this approach has provided a wealth of information about numerous helicases^{15,17–20}, it requires laborious synthesis of multiple substrates, lacks single-nucleotide resolution and the results are potentially influenced by end-effects at the duplex termini. Recent advances in single molecule technology have provided information about overall unwinding rate constants and processivity^{21–24}, however, they lack the resolution to report on microscopic behaviour at the single nucleotide level.

To address these issues and to build a mechanistic framework for



RNA unwinding by NS3, we developed a combinatorial method for monitoring helicase behaviour at each nucleotide along a duplex substrate (Fig. 1). A large family of multiple-length substrates was created by doping RNA transcripts with phosphorothioate nucleoside 5'-triphosphates (NTP's). The resultant pool of transcripts were 5'-end labelled and annealed to a complementary 'bottom strand' (which contained a 3'-overhang, as required by the NS3 helicase). The phosphorothioate linkages were then cleaved with iodine to generate a large family of substrates that contained, on average, a single nick at each position along the top strand (randomly nicked substrates, RNS). The resultant RNS pool was incubated with helicase in order to build up a population of functional complexes at the junction, and then synchronized unwinding was initiated by the addition of ATP and a trap oligonucleotide ('single-cycle conditions', see Methods). Microscopic rate constants and processivity at each nucleotide were quantitated by analysis of individual bands on a sequencing gel.

The simplest application of this experiment involved collecting information on the extent of unwinding at a single point in time. For example, in one set of experiments, NS3 was bound to an RNS duplex of 80 base pairs (bp), the reaction was initiated with ATP/trap oligonucleotide and allowed to proceed briefly (30 s) before loading onto a single gel lane. After quantitation, the fraction of unwound duplex was plotted as a function of duplex length (Fig. 2a). The resultant raw data displayed a clear stepping pattern that represented phased patterns of behaviour at distinct regions of the duplex. After taking the derivative of the smoothed data, we observed a sinusoidal pattern of peaks and valleys that reflected the relative helicase processivity at each nucleotide position along the duplex. These data indicate that, after NS3 initiates at the single-stranded junction (position 0 on the x-axis of Fig. 2b), there is a

periodic change in unwinding behaviour that occurs approximately every 18 bps (the 'physical step'). Peaks and valleys in the data indicate that between steps (the valleys), the helicase unwinds with high processivity (that is, it proceeds directly and does not fall off), but at step boundaries (the peaks), the helicase tends to dissociate from the substrate.

In order to develop a mechanistic understanding of the observed 'physical step' and to characterize kinetic events, RNS unwinding by NS3 was monitored as a function of time. Time courses were performed in a quench-flow apparatus, loaded on successive lanes of a high-resolution native gel (Fig. 1, right panel), and the resulting data were analysed in a manner analogous to that developed by Lohman^{17,18}, in which time courses for substrates of increasing length were subjected to global fitting in order to determine the apparent kinetic step size and the individual rate constants. In this case, however, we can interrogate a potentially infinite number of substrates at each nucleotide, analyse their behaviour simultaneously and the resultant data are uncompromised by divergent behaviour as the helicase approaches a duplex terminus. The amplitude data at single time points (Fig. 2) and the kinetic data (Fig. 3) were highly reproducible, having been repeated at least

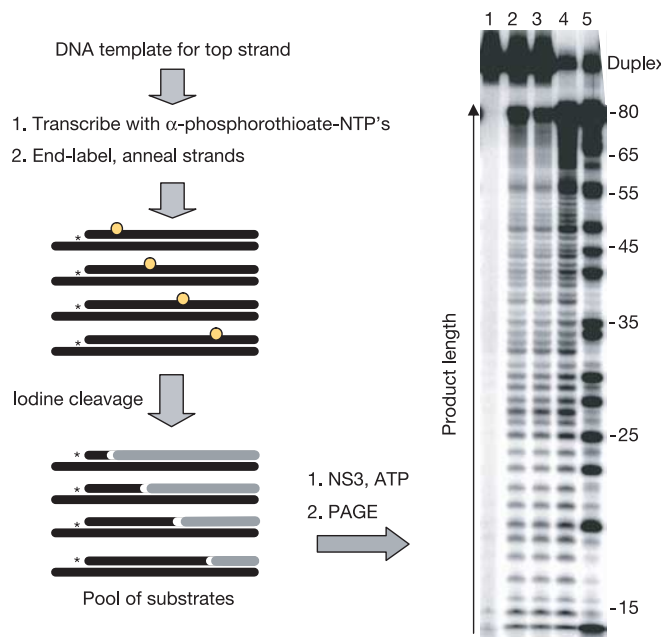


Figure 1 Schematic diagram of the RNS unwinding assay and subsequent product analysis. (Left) A pool of RNA duplexes containing random phosphorothioate substitutions. Positions of phosphorothioate linkages (yellow dots), and the 5'-terminal ³²P on the top strands (asterisks). After iodine treatment, the resultant nick is represented by the gap between pieces of top-strand RNA (black and grey). (Right) Semi-native sequencing gel (15%) showing products of NS3 reaction with an 80-bp duplex without ATP (lane 1), and with ATP for 30 and 60 s (Lanes 2 and 3, respectively). The heat-denatured RNA control (lane 4, 98% denatured) is shown next to a uridine marker lane (lane 5; α -phosphorothioate-UTP doped, I₂-treated). Product positions are marked at right. To obtain single-nucleotide resolution of higher-molecular weight products (55–80 nucleotides), samples were also run on a 12% gel (not shown).

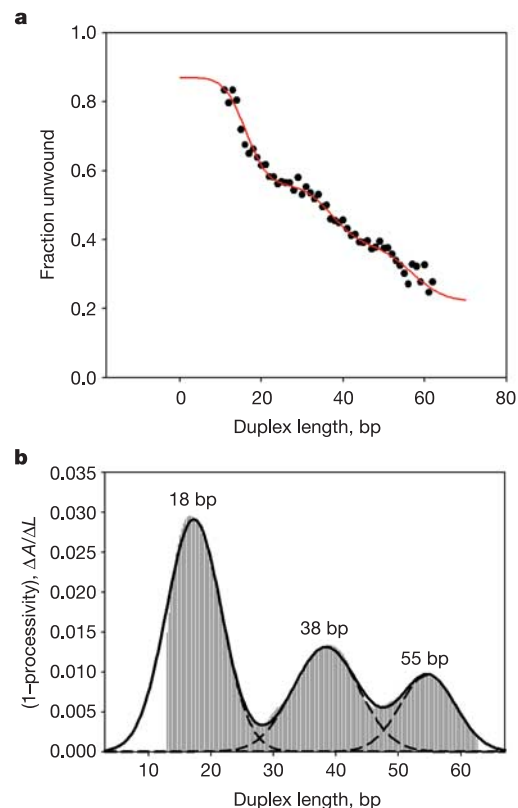


Figure 2 Determination of the NS3 physical step size. **a**, Unwinding amplitudes for RNS duplexes of varying length ($t = 30$ s). Fraction unwound at 30 s is computed as the ratio of c.p.m. in each product band (lane 2, Fig. 1) relative to the corresponding band in the heat-denatured control (lane 4, Fig. 1). Each data point is the average of four independent experiments. The solid line represents the integral of the gaussian fit obtained in Fig. 2b. **b**, Distribution of local processivity versus duplex length. The raw data from **a** were smoothed with a loess algorithm. After taking the derivative of the smoothed data (shaded peaks, $\Delta A/\Delta L$, where A is the unwinding amplitude from **a** and L is the corresponding RNS duplex length), the resultant curves were best fit by a function that represents the simple sum of three gaussians (solid line), resulting in peak positions of 18.2 ± 0.04 bp (mean $\pm$ standard error), 38.4 ± 0.07 bp and 54.7 ± 0.1 bp. Experiments were conducted at three different protein concentrations: 120 nM (shown), 70 nM and 160 nM, with no difference in the resulting parameters. Experiments were also conducted at three ATP concentrations: 4 mM (shown), 0.4 mM and 0.1 mM. The only observed effect was a sequential increase in peak amplitude as [ATP] was lowered (not shown).

four times using two different types of substrates (one containing a 60-nucleotide duplex, and another using an 80-nucleotide duplex).

Representative time courses showed four types of kinetic behaviour during the unwinding of the 80-mer duplex (Fig. 3). Time courses from within the first twenty bp were highly superimposable and they could be described by the same kinetic model with a single

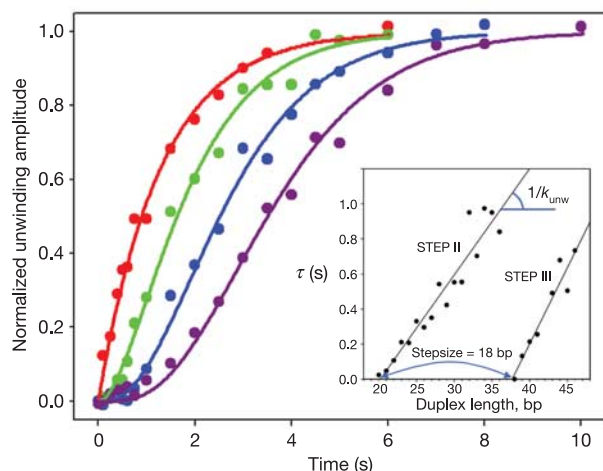


Figure 3 NS3 unwinding kinetics as a function of duplex length. Representative time courses are shown for duplexes belonging to each of the four mechanistic classes observed in this study. All kinetic curves fall into one of these four patterns, depending on duplex length. Duplex lengths up to 19 bp (17 bp shown) (red), duplex lengths 20–38 (30 bp shown) (green), duplex lengths 39–57 (47 bp shown) (blue), and duplex lengths from 58–76 (65 bp shown) (purple). Solid lines represent the best fit using one, two, three or four major kinetic steps, respectively. Analysis of the microscopic unwinding rate constant (Inset). The time required (τ) for unwinding of bp between paused states is directly proportional to the duplex length. The slope of the resultant line is equal to $1/k_{\text{unw}}$. Spacing between major steps II and III provides another measure of kinetic step size (18 bp).

rate limiting event (Fig. 3, red curve). Abruptly, this behaviour changed and bp 21–40 displayed a different kinetic behaviour that was reflected in a larger lag phase and which required a second major rate limiting event (Fig. 3, green curve). Similarly, bp 41–60 (blue curve) and 61–80 (purple curve) required three and four rate limiting events, respectively. These effects indicate that NS3 unwinds the duplex with a kinetic step size of approximately 20 bp. Global fitting of the data using conventional methods (see Supplementary Methods; ref. 18) yielded an overall kinetic rate constant of 0.8 s^{-1} per step. Agreement among the overall rate constants indicates that each successive periodic event, or step, is identical except for the first one, which seems to be limited by initiation (0.4 s^{-1} , see Supplementary Methods). Because the kinetic step size was derived from the magnitude of the lag in time courses¹⁸, and not from the relative changes in reaction extent (amplitude), the kinetic data yielded an independent measure of ‘step size’ that was directly comparable with estimates of the ‘physical step’ that were demonstrated through amplitude effects (as in Fig. 2b). It is therefore significant that the two values agree.

In addition to the major rate-limiting event that occurs after ~ 20 bp of unwinding, the data were sufficiently resolved to describe events that occur within the kinetic step itself. If unwinding within the step were infinitely rapid, then time courses from within the same step would be perfectly superimposable. But there were, in fact, small deviations (Supplementary Information): although the curves fit the same overall kinetic model (that is, they are described by the same number of major rate-limiting events), there was a small, progressive decrease in unwinding efficiency for duplexes of increasing length, which reflected the increased time (τ) required for NS3 to unwind a longer duplex (Fig. 3, inset). From this decrease in observed rate constant, it was possible to calculate the microscopic rate constant for unwinding bp within a given kinetic/physical step (k_{unw} , 16.5 bp s^{-1} ; Fig. 3, inset), to differentiate this translocation rate constant from the slow, local event that occurs at step boundaries (k_{p} , 5.6 s^{-1} , see Supplementary Information) and to obtain a more precise estimate of the kinetic step size (18 bp;

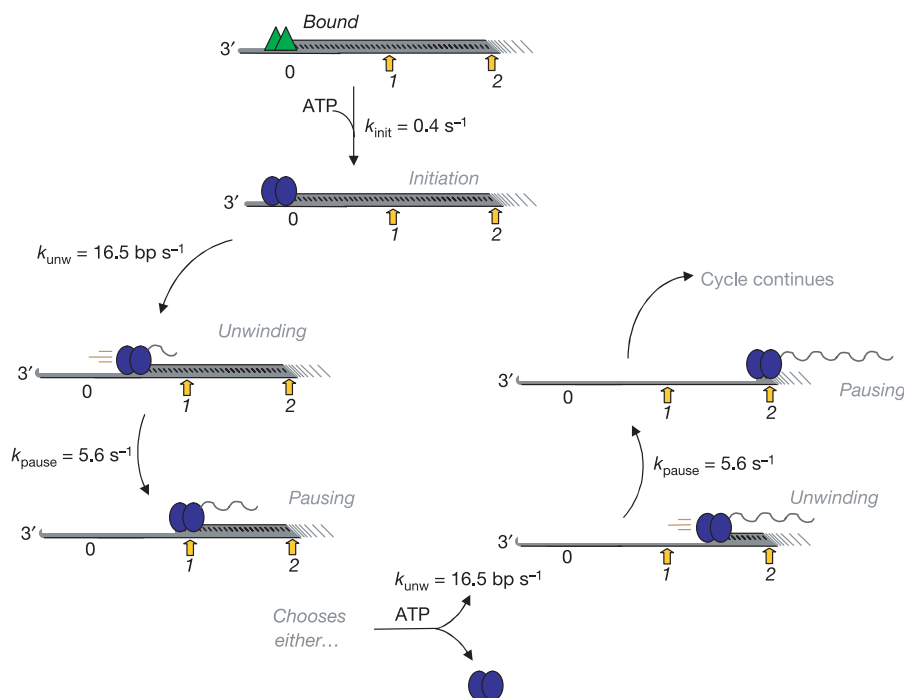


Figure 4 The kinetic scheme for periodic NS3 unwinding of RNA. The NS3 is represented as a dimer (see Supplementary Information) in activated (blue elipsoids) and inactivated (green triangles) states. Unwinding (unw), initiation (init).

Fig. 3, inset). This analysis of microscopic events demonstrated a series of coordinated, discontinuous movements that were required for unwinding RNA by the NS3 helicase. Taken together, the data are consistent with a coherent kinetic mechanism for NS3 unwinding of RNA (Fig. 4).

Three of the most interesting questions raised by these observations are: (1) What is happening at the boundaries between 18 bp increments? (2) What is the meaning of the 18 bp kinetic and physical step? (3) How is helicase behaviour influenced by sequence? Some insight into the first question is provided by the ATP dependence of the reaction. When the ATP concentration is altered, the only parameter that was affected was the reaction amplitude at the step boundaries (that is, the height of the peaks in Fig. 2b, data not shown). This indicates that, as with helicase NPH-II²⁰, ATP binding has a direct role in regulating processivity of the enzyme. In this case, however, experimental resolution was sufficient to show that the effect was exerted specifically at the step boundaries. Thus, the boundary state is likely to represent a position at which NS3 adjusts its conformation in an ATP-dependent manner. These findings do not rule out the involvement of ATP at other points in the reaction cycle, such as during translocation.

To address the second question, it is useful to consider that the physical and kinetic step size of NS3 is relatively large by comparison with other SF2 and SF1 helicases, which have been reported to have shorter step sizes of 4–6 bp^{18,20,25}. However, the heteromeric RecBC helicase takes a step size of 23 bp¹⁹. Based on these observations, we reasoned that the observed step size might be related both to the specific reaction mechanism and the molecular size of the functional helicase complex. To address this issue, we evaluated the oligomeric state of the NS3 protein. Mutational and biochemical studies have long suggested that NS3 behaves as a multimer, and more specifically, it has been proposed to form a dimer^{2,10,26,27}. Although we observed that monomeric NS3 was capable of binding RNA with high affinity, active RNA unwinding and functional complex formation required a dimeric form of NS3 (see Supplementary Information, also Fig. 4).

The observed step size is insensitive to substrate tail length (Supplementary Information) and ATP concentration, which suggests that it is a fundamental property of NS3 unwinding. However, it remains possible that the unwinding parameters will be influenced by sequence content and that NS3 can report on local alterations in thermodynamic stability. Furthermore, it is likely that step size can be influenced by the physico-chemical properties of the tracking (bottom) strand. For example, a kinetic step size of 9.1 ± 0.9 bp was recently reported for DNA unwinding by NS3 (ref. 28). This is consistent with the fact DNA and RNA unwinding by NS3 display numerous mechanistic differences¹⁵ and that the behaviour of SF2 helicases is sensitive to the backbone composition and conformation of the tracking strand²⁹. Whereas NS3 may be a multifunctional motor with several strategies for nucleic acid unwinding, the data herein demonstrate that helicases (much like cytoskeletal motor proteins)¹² can move discontinuously through a physically defined, kinetic cycle as they translocate along a polymer and execute mechanical work. □

Methods

Preparation of RNS duplexes

RNA strands were prepared by *in vitro* transcription of polymerase chain reaction (PCR)-generated DNA templates using the MEGAscript T7 RNA transcription kit (Ambion). To generate a pool of top-strand oligonucleotides that were doped with phosphorothioates, four separate transcription reactions were conducted for each different type of α -phosphorothioate NTP (the polymerase was less efficient when more than one type of α -phosphorothioate NTP was present in the reaction mix). In the transcription reactions, the concentration of α -phosphorothioate NTP was 5 times lower (0.5 mM) than that of the NTPs (2.5 mM), resulting in strands that contained approximately 0.3 modifications per oligonucleotide (thus ensuring against double modifications). After transcription, the four different types of strands (that is, polymers

doped with α -phosphorothioate ATP, CTP, GTP and UTP) were combined in equimolar proportions to yield a single pool of top-strand molecules that contain randomly incorporated phosphorothioate linkages at low frequency.

Top strand oligonucleotides were either 60 or 80 nucleotides in length, as indicated. The 80-mer sequence was 5'-GGGAGACUACGUUCGACUAGUGUACUCUGCCUUGCGACUGCUGGCCUCCAGGUCUCCCGAACAGUCCUCACACUCCC-3'. The 60-mer sequence was identical, although 20 nucleotides shorter at the 3'-end. Bottom strand RNAs were 18 nucleotides longer than the complementary top strands so that, on annealing with the respective top strand, each substrate contained a 60- or 80-bp duplex that was joined to an 18 nucleotide single-stranded 3'-overhang on the bottom strand. The overhang sequence was 5'-AGGAGCUCAGCUAUCAGA-3'. As illustrated (Fig. 1), bottom strand RNAs were unmodified.

To prepare RNS duplexes for reaction, a pool of top-strand oligonucleotides was 5'-end labelled with [γ -³²P]ATP and annealed with 2–4 $\times$ molar excess of unlabelled bottom strand in 10 mM MOPS-NH₄OH at pH 6.5, 30 mM NaCl and 1 mM EDTA as previously described²⁰. The phosphorothioate linkages were then cleaved by addition of iodine (to 0.1 mM final concentration) and an incubation for 15 min at room temperature. The RNS duplex pool was then quenched and purified by loading on a partially native polyacrylamide gel (10% polyacrylamide, 3 M urea, 0.5 $\times$ Tris-borate electrophoresis buffer, TBE).

Unwinding reactions

Unwinding reactions were performed at 37 °C in a buffer containing 30 mM MOPS-NH₄OH at pH 6.5, 30 mM NaCl, 3 mM MgCl₂, 1% glycerol, 1–2 nM RNA substrate and 100–160 nM NS3 protein. RNS duplex was first incubated with NS3 in reaction buffer for 1 h to promote functional association¹⁵. The unwinding reaction was then initiated by simultaneous addition of ATP (4 mM) and trap oligonucleotide (2 μ M). The trap oligonucleotide, which was previously demonstrated to be a high affinity, rapidly binding ligand for NS3¹⁵, is a 60-mer DNA oligonucleotide with a sequence unrelated to the RNS. For amplitude studies (as in Fig. 2), the reaction was quenched at a specific time point (for example, 30 or 60 s) by adding 0.25 volumes of 5 $\times$ loading buffer (25 mM EDTA, 1% SDS, 50% sucrose, 0.05% bromophenol blue and 0.05% xylene cyanol). The products of unwinding were resolved on a semi-native gel (15% acrylamide, 3 M urea and 0.5 $\times$ TBE) and bands corresponding to the released top strands were quantitated by a phosphorimager. The fraction unwound for each top-strand length was determined by assuming that the corresponding band for a boiled substrate (completely unwound) represents 100% unwinding. Although urea was introduced to improve resolution of the bands on the gel, it did not promote dissociation of the duplex substrate fractions. Owing to the extremely low concentrations of RNA in the unwinding experiments, reannealing of dissociated RNA strands did not occur.

Chemical quenched flow experiments were performed using model RQF-3 Kintek rapid quench-flow apparatus (Kintek) maintained at 37 °C using a circulating water bath. The unwinding reactions were quenched at various times (ranging from 5 ms to 60 s) and the products were immediately analysed by polyacrylamide gel electrophoresis (PAGE) 15% polyacrylamide/3 M urea/0.5 $\times$ TBE.

In all cases bands were visualized and quantitated using a Storm PhosphorImager and ImageQuant software (Molecular Dynamics).

Kinetic analysis

Kinetic analysis of unwinding time courses was performed using the global fitting program DYNAFIT (BioKin)³⁰. All kinetic time courses (42 time courses for the 60 bp substrate and 55 time courses for the 80 bp substrate) were subjected to global fitting by applying the following kinetic scheme:



S and P are the substrate and fully unwound duplexes, respectively, P_n is the paused state and T_n is the translocating state during the *n*th step. All practical details regarding the fitting are provided in Supplementary Methods.

Received 13 February; accepted 1 June 2004; doi:10.1038/nature02704.

- Moradpour, D. *et al.* Membrane association of hepatitis C virus nonstructural proteins and identification of the membrane alteration that harbors the viral replication complex. *Antiviral Res.* **60**, 103–109 (2003).
- Dimitrova, M., Imbert, I., Kieny, M. P. & Schuster, C. Protein–protein interactions between hepatitis C virus nonstructural proteins. *J. Virol.* **77**, 5401–5414 (2003).
- Piccinini, S. *et al.* Modulation of the hepatitis C virus RNA-dependent RNA polymerase activity by the non-structural (NS) 3 helicase and the NS4B membrane protein. *J. Biol. Chem.* **277**, 45670–45679 (2002).
- Gallinari, P. *et al.* Modulation of hepatitis C virus NS3 protease and helicase activities through the interaction with NS4A. *Biochemistry* **38**, 5620–5632 (1999).
- Bird, L. E., Subramanya, H. S. & Wigley, D. B. Helicases: a unifying structural theme? *Curr. Opin. Struct. Biol.* **8**, 14–18 (1998).
- Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326 (1998).
- Caruthers, J. M. & McKay, D. B. Helicase structure and mechanism. *Curr. Opin. Struct. Biol.* **12**, 123–133 (2002).
- Kim, J. L. *et al.* Hepatitis C virus NS3 RNA helicase domain with a bound oligonucleotide: the crystal structure provides insights into the mode of unwinding. *Structure* **6**, 89–100 (1998).
- Kim, D. W., Kim, J., Gwack, Y., Han, J. H. & Choe, J. Mutational analysis of the hepatitis C virus RNA helicase. *J. Virol.* **71**, 9400–9409 (1997).
- Cho, H.-S. *et al.* Crystal structure of RNA helicase from genotype 1b hepatitis C virus. A feasible mechanism of unwinding duplex RNA. *J. Biol. Chem.* **273**, 15045–15052 (1998).
- Korolev, S., Yao, N., Lohman, T. M., Weber, P. C. & Waksman, G. Comparisons between the structures

- of HCV and Rep helicases reveal structural similarities between SF1 and SF2 super-families of helicases. *Protein Sci.* **7**, 605–610 (1998).
12. Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288**, 88–95 (2000).
 13. Bartenschlager, R. The NS3/4A proteinase of the hepatitis C virus: unravelling structure and function of an unusual enzyme and a prime target for antiviral therapy. *J. Viral Hepat.* **6**, 165–181 (1999).
 14. Gallinari, P. *et al.* Multiple enzymatic activities associated with recombinant NS3 protein of hepatitis C virus. *J. Virol.* **72**, 6758–6769 (1998).
 15. Pang, P. S., Jankowsky, E., Planet, P. J. & Pyle, A. M. The hepatitis C viral NS3 protein is a processive DNA helicase with cofactor enhanced RNA unwinding. *EMBO J.* **21**, 1168–1176 (2002).
 16. Paolini, C., De Francesco, R. & Gallinari, P. Enzymatic properties of hepatitis C virus NS3-associated helicase. *J. Gen. Virol.* **81**, 1335–1345 (2000).
 17. Lucius, A. L., Maluf, N. K., Fischer, C. J. & Lohman, T. M. General methods for analysis of sequential "n-step" kinetic mechanisms: application to single turnover kinetics of helicase-catalyzed DNA unwinding. *Biophys. J.* **85**, 2224–2239 (2003).
 18. Ali, J. A. & Lohman, T. M. Kinetic measurement of the step size of DNA unwinding by *Escherichia coli* UvrD helicase. *Science* **275**, 377–380 (1997).
 19. Bianco, P. R. & Kowalczykowski, S. C. Translocation step size and mechanism of the RecBC DNA helicase. *Nature* **405**, 368–372 (2000).
 20. Jankowsky, E., Gross, C. H., Shuman, S. & Pyle, A. M. The DExH protein NPH-II is a processive and directional motor for unwinding RNA. *Nature* **403**, 447–451 (2000).
 21. Bianco, P. R. *et al.* Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. *Nature* **409**, 374–378 (2001).
 22. Ha, T. *et al.* Initiation and re-initiation of DNA unwinding by the *Escherichia coli* Rep helicase. *Nature* **419**, 638–641 (2002).
 23. Spies, M. *et al.* A molecular ratchet: the recombination hotspot chi controls DNA translocation by the RecBCD helicase. *Cell* **114**, 647–654 (2003).
 24. Dessinges, M. N., Lionnet, T., Xi, X. G., Bensimon, D. & Croquette, V. Single-molecule assay reveals strand switching and enhanced processivity of UvrD. *Proc. Natl Acad. Sci. USA* **101**, 6439–6444 (2004).
 25. Lucius, A. L. *et al.* DNA unwinding step-size of *E. coli* RecBCD helicase determined from single turnover chemical quenched-flow kinetic studies. *J. Mol. Biol.* **324**, 409–428 (2002).
 26. Levin, M. K. & Patel, S. S. The helicase from hepatitis C virus is active as an oligomer. *J. Biol. Chem.* **274**, 31839–31846 (1999).
 27. Locatelli, G. A., Spadari, S. & Maga, G. Hepatitis C virus NS3 ATPase/helicase: an ATP switch regulates the cooperativity among the different substrate binding sites. *Biochemistry* **41**, 10332–10342 (2002).
 28. Levin, M. K., Wang, Y. H. & Patel, S. S. The functional interaction of the hepatitis C virus helicase molecules is responsible for unwinding processivity. *J. Biol. Chem.* **279**, 26005–26012 (2004).
 29. Kawaoka, J., Jankowsky, E. & Pyle, A. M. Backbone tracking by the SF2 helicase NPH-II. *Nature Struct. Mol. Biol.* **11**, 526–530 (2004).
 30. Kuzmic, P. Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase. *Anal. Biochem.* **237**, 260–273 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The authors would like to thank E. Jankowsky, J. Kawaoka and M. Brenowitz for discussions. We thank R. Beran for sharing his results on NS3 unwinding of multi-piece RNA substrates, and H. Le for discussions and early gifts of NS3 protein. This work was supported by grants from the NIH and the Howard Hughes Medical Institute. A.M.P. is an HHMI investigator.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.M.P. (anna.pyle@yale.edu).

Detection of an intermediate of photosynthetic water oxidation

Juergen Clausen & Wolfgang Junge

Division of Biophysics, Department of Biology/Chemistry, Universität Osnabrück, D-49069 Osnabrück, Germany

The oxygen that we breathe is produced by photosystem II of cyanobacteria and plants. The catalytic centre, a Mn_4Ca cluster, accumulates four oxidizing equivalents before oxygen is formed, seemingly in a single reaction step^{1–8} $2\text{H}_2\text{O} \rightleftharpoons \text{O}_2 + 4\text{H}^+ + 4\text{e}^-$. The energy and cycling of this reaction derives solely from light. No intermediate oxidation product of water has been detected so far. Here, we shifted the equilibrium of the terminal reaction

backward by increasing the oxygen pressure and monitoring (by absorption transients in the near-ultraviolet spectrum) the electron transfer from bound water into the catalytic centre. A tenfold increase of ambient oxygen pressure (2.3 bar) half-suppressed the full progression to oxygen. The remaining electron transfer at saturating pressure (30 bar) was compatible with the formation of a stabilized intermediate. The abstraction of four electrons from water was probably split into at least two electron transfers: mildly endergonic from the centre's highest oxidation state to an intermediate, and exergonic from the intermediate to oxygen. There is little leeway for photosynthetic organisms to push the atmospheric oxygen concentration much above the present level.

The catalytic centre of photosynthetic water oxidation is embedded in a protein, photosystem II (PSII), and is composed of a Mn_4Ca cluster and a redox-active tyrosine (Y_Z)^{1–3}. Its structure is beginning to be clarified^{4–6}. Absorption of a quantum of light drives a charge separation to yield an oxidized chlorophyll *a* entity (P680^+). It oxidizes Y_Z in nanoseconds and, from there on, the Mn_4Ca cluster in microseconds^{2,7}. Three absorbed quanta clock the Mn_4Ca cluster through increasingly higher oxidation states, from the lowest, S_0 , to S_1 , S_2 and S_3 . Dioxygen is liberated only after input of a fourth quantum of light to yield $\text{S}_3\text{Y}_Z^{\text{ox}} = \text{S}_4^8$. Up to the penultimate state (S_3) the substrate, bound water, is exchangeable with isotope-labelled bulk water^{9,10}. The production of potentially dangerous oxidation intermediates of water seems to be avoided by concentrating the fourfold oxidation of two bound water molecules into a single terminal reaction, $\text{S}_4(\text{H}_2\text{O})_2 \rightarrow \text{S}_0\text{Y}_Z + \text{O}_2 + \text{protons}$. Despite intensive search, oxidized intermediates of water (for example, the postulated peroxide¹¹) have escaped detection¹². There are three factors indicative of the terminal reaction sequence, from start to end, which evolve with the same velocity: (1) the reduction of Y_Z^{ox} as detected by electron paramagnetic resonance¹³; (2) the reduction of Mn_4CaY_Z (by bound water), as detected by ultraviolet flash spectrophotometry^{14–17}; and (3) the liberation of oxygen, as detected by rapid polarography^{12,17,18}. The terminal reaction sequence is probably rate-limited by the transfer of the first electron into $\text{S}_3\text{Y}_Z^{\text{ox}}$. Intermediates might have escaped detection because of their short lifetime.

We attempted to stabilize intermediates by increasing the partial pressure of oxygen over the reaction mixture. This approach is illustrated by a sequential reaction with only one

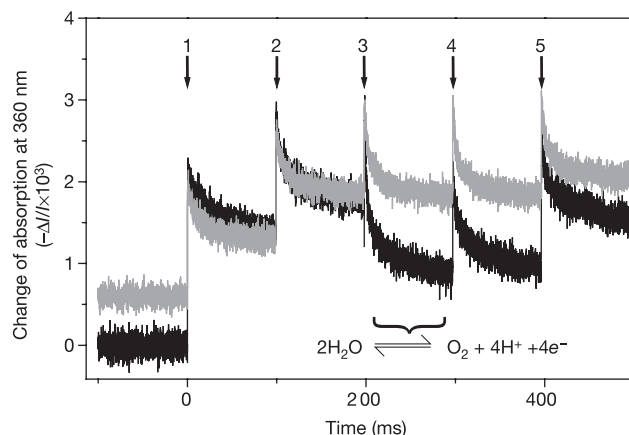
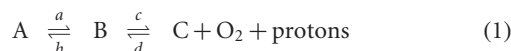
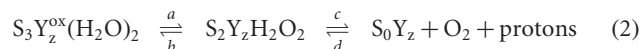


Figure 1 Ultraviolet absorption transients at a wavelength of 360 nm of dark-adapted PSII core particles under excitation with five short laser flashes (arrows). Black trace, air; grey trace, 20 bar oxygen. The undershoot after the third flash (black trace) reflects the relaxation of the fourfold oxidized catalytic centre culminating in the release of oxygen (see text).

intermediate, B:



which may be tentatively read as:



$a - d$ denote the forward and backward rate constants, and H_2O_2 stands for the hypothetical peroxide intermediate. The choice of only a single intermediate and its assignment to S_2 -peroxide is arbitrary but plausible. All reactions are pseudo-first order, except for the last back reaction (second-order rate constant d). With S_i representing solid state components, and water being both reagent and solvent, oxygen and protons are the only dilute reaction partners. The involvement of protons was neglected in the following for simplicity. We varied the oxygen concentration to change the equilibrium between the intermediate (B) and the fully relaxed state (C).

Figure 1 shows absorption transients at a wavelength of 360 nm under excitation of dark-adapted PSII core particles with a train of five laser flashes. These transients reflect redox reactions of manganese and of intrinsic and extrinsic electron acceptors of PSII, as documented in the literature^{17,19} (see also Supplementary Fig. 2). Both traces were recorded at pH 6.7, the first at ambient oxygen pressure (black) and the second at 20 bar oxygen (grey). The most interesting feature of the control trace (black) is the decrease in the absorption transient upon the third flash of light, which hereafter we call the undershoot. It is caused by an exponential decay component with a half-life of typically 1.3 ms (20 °C). It compensates the upward-directed jumps at flashes 1 and 2 (see Supplementary Information) and is directly related to oxygen evolution^{12,14–17,20}. In PSII particles from *Synechocystis*, as used in this work, the equilibrium of the oxidation states in the dark is $0.3/0.7/0/0 = S_0/S_1/S_2/S_3$ (refs 12, 21). Thus, oxygen evolution and the concomitant undershoot of absorption appear only at and after the third flash of light, and are absent at the former two flashes^{17,21}. Comparison of the two traces in Fig. 1 reveals that the undershoot is suppressed at high oxygen pressure (grey transient).

There are more differences between the two traces in Fig. 1: (1) the magnitude of the upshot at the first flash differed because of a different proportion of incompetent centres in oxygen evolution (see Supplementary Information), and (2) the non-oscillating signals at flashes 4 and 5, and further on, at high oxygen pressure.

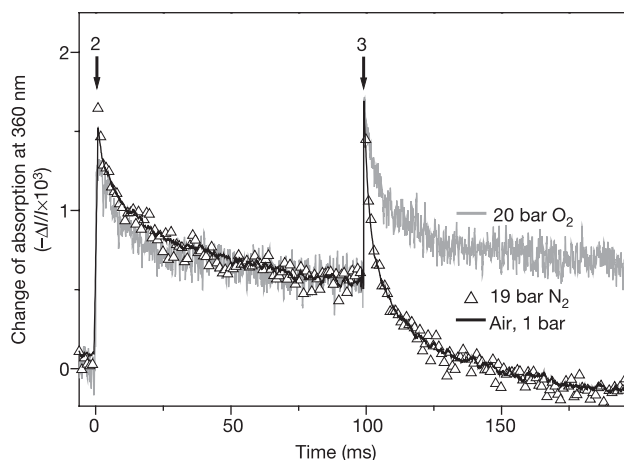


Figure 2 Ultraviolet absorption transients as induced by the second and third flash under variation of the gas composition and pressure. The undershoot indicating the formation of oxygen is clearly depressed by 20 bar O_2 but is unaffected by enhanced hydrostatic pressure alone (19 bar N_2). The control trace (air) is smoothed over 30 points, whereas the other traces are smoothed over five points.

We tentatively interpreted the latter as a cycling of <100 ms between states S_3Y^{ox} and S_3Y at the donor side of PSII, similar to that which occurs in Mn-depleted PSII core particles.

Figure 2 shows the original transients of flashes 2 and 3 under three conditions: with ambient oxygen pressure (0.21 bar); with elevated oxygen pressure (20 bar); and with elevated nitrogen pressure and no oxygen present (19 bar N_2 , only a few smoothed points displayed). It is evident that the undershoot was diminished by elevated oxygen pressure, but not by pressure *per se*. In contrast to the decay after the third flash, the decay after the second flash was not affected by the elevated oxygen pressure, implying that the dark equilibrium between oxidation states was not perturbed. The effects caused by exposure to 30 bar oxygen were reversible. After 30 min exposure to 20 bar O_2 and slow re-establishment of the ambient pressure (over 5 h at 0 °C), samples still revealed typically 70–80% of their original oxygen-evolving activity under continuous illumination. Similarly, the normal oscillatory behaviour under excitation of dark-adapted material with a group of light flashes was restored by decompression to the ambient oxygen level. This is documented in the Supplementary Information.

The effects of increased oxygen pressure were immediately apparent on visual inspection of the raw data. The ultraviolet transients at the third laser flash were normalized to yield the same extent of the rapid rise, and corrected (see Methods and Fig. 3). The normalized and corrected transients at 360 nm at the third laser flash reflect the electron transfer into the catalytic centre after reaching state S_4 . Three corrected traces (at 0.21, 1 and 20 bar O_2) are documented in Fig. 3. We titrated their extent as a function of the oxygen pressure (see Fig. 4, points). The extrapolated saturation level of the suppression by high oxygen pressure was 83% (dashed line) and the midpoint was at 2.3 bar.

The solid line in Fig. 4 was calculated on the basis of the tentatively chosen reaction scheme (equation (1)) involving only a single intermediate. Initially (that is, right after the third laser flash) the concentrations of A, B and C were $A_0 = 1$ and $B_0 = C_0 = 0$, and the new equilibrium was:

$$A_{\infty} = \left(1 + \frac{a}{b} \left(1 + \frac{c}{pd}\right)\right)^{-1}, \quad \frac{B_{\infty}}{A_{\infty}} = \frac{a}{b}, \quad \frac{C_{\infty}}{B_{\infty}} = \frac{c}{pd} \quad (3)$$

where p denotes the oxygen pressure and a , b , c and d the rate constant of the reaction scheme (equation (1)). The three species A,

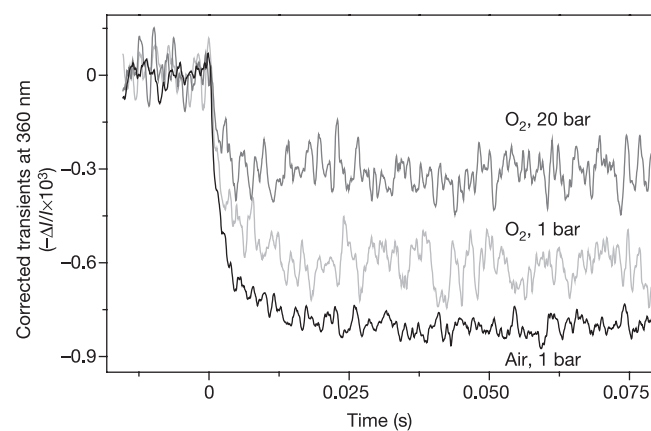


Figure 3 Electron transfer into the catalytic Mn_4Ca centre at three different oxygen pressures, and corrected ultraviolet absorption transients at the third flash. Each data point represents the difference between the respective points at the third and the second flash (as shown in Fig. 2), serving to eliminate the contribution at the chosen wavelength (360 nm) of the consecutive reactions of the added electron acceptor (2,5-dichloro-*p*-benzoquinone). A further correction was introduced for the underlying step-shaped signal attributable to the transition $S_2 \rightarrow S_3$ on the second flash. Original data were smoothed over 20 points.

B and C could not be monitored separately, and together they contributed to the corrected ultraviolet transient at the third flash of light. We discuss the expected behaviour in terms of equation (2), wherein the intermediate is assumed to be $S_2Y_ZH_2O_2$. This is only one of several choices, as mentioned. The decay of absorption reverses two positive jumps of equal magnitude during the foregoing univalent oxidation steps, $S_1 \rightarrow S_2$ and $S_2 \rightarrow S_3$ (refs 22–24) (see Supplementary Information). In contrast to the former, the transitions $Y_Z \rightarrow Y_Z^{ox}$ and $S_0 \rightarrow S_1$ marginally contribute at this wavelength (360 nm)²². Accordingly, the disappearance of $A = S_3 \cdot Y_Z^{ox} (H_2O)_2$, the highest oxidation state of the centre, is expected to cause absorption twice as large as the same amount of an intermediate in the second oxidation state, $B = S_2Y_ZH_2O_2$.

We interpreted the corrected ultraviolet transient, therefore, as reflecting the behaviour of the composite observable, X , as a function of the oxygen pressure (p):

$$X = A + \frac{1}{2}B = f(p) \quad (4)$$

Insertion of equation (3) into equation (4) yielded the documented fit (solid line) to the titration data in Fig. 4. The fit parameter values were $a/b = 0.5$, $c/d = 6.8$ bar and $p_{1/2} = 2.3$ bar.

The expected time course at very high oxygen pressure ($p \rightarrow \infty$) is given by:

$$X(t) = \frac{1}{2} \left(1 + \frac{1}{a+b} (b + ae^{-(a+b)t}) \right) \quad (5)$$

It is a mono-exponential decay with a rate constant, k :

$$k = a + b \quad (6)$$

At 20 bar O_2 the observed half-decay time of the signal was 0.8 ms (data fit by single exponential), implying a rate constant of $k = 866 s^{-1}$, so that $a = 144 s^{-1}$ and $b = 722 s^{-1}$. The absolute values of the rate constants of the second partial reaction, c and dp , could not be determined, but both were much larger than a and b .

With these parameters, which were based on data obtained at pH 6.7 and at 20 °C, the standard free energy profile of the reaction sequence (equation (1)) could be calculated (see equation (3)) according to the relation $\Delta G^0 = -RT \ln K$, where RT is as usual and K denotes the equilibrium constant.

$$\Delta G_{AB}^{0'} \cong +1.7 kJ mol^{-1}, \quad \Delta G_{BC}^{0'} \cong -4.7 kJ mol^{-1} \quad (7)$$

Both values relate to the production of 1 mol dioxygen under pseudo-standard conditions ($\approx$ pH 6.7, 1 bar O_2 , 20 °C). This free

energy profile is illustrated in the insert to Fig. 4. At ambient pressure (0.21 bar oxygen) ΔG_{BC} would be $-8.6 kJ mol^{-1}$. The role of proton release in the final reaction sequence will be detailed elsewhere. Here, we only mention that the second partial reaction ($B \rightarrow C$) involves the liberation into the bulk of one proton per O_2 . This will be detailed elsewhere (see refs 17, 25–28 for surveys of the proteolytic reactions).

Our experiments demonstrate the back pressure of oxygen on the Mn_4Ca centre. It would be interesting to re-investigate the previously reported oxygen consumption by PSII²⁹ (under ambient pressure) under conditions of varied oxygen pressure.

Production of oxygen through photosynthetic water oxidation is a process of paramount biological importance and technical relevance. It operates at extremely positive redox potential ($>+1.2$ V; refs 25, 30). We present experimental evidence for an intermediate of the enigmatic reaction pathway from water to oxygen. The free energy drop at the very end of this reaction is rather small. It emphasizes that PSII operates at the outermost redox span that can be driven by quanta of red light. The unexpectedly low pressure that half-suppresses oxygen evolution (2.3 bar at pH 6.7) has an interesting consequence. The present atmospheric pressure of oxygen is close to the expected obtainable maximum. It approaches it rather closely when taking into consideration the additional effect of pH: the luminal side of the thylakoid membrane is more acidic than the pH 6.7 used in our experiments. □

Methods

PSII core particles and media

Oxygen-evolving PSII core particles were prepared as described²¹ from modified WT* cells of *Synechocystis* sp. PCC6803 (ref. 21). For flash-spectrometric measurements PSII core particles were suspended in standard medium (1 M sucrose, 25 mM $CaCl_2$, 10 mM NaCl, 1 M glycine betaine, 0.06% (w/v) *N*-Dodecyl- β -D-maltoside, 50 mM 2-(*N*-Morpholino)ethanesulphonic acid, pH 6.7). The temperature was 20 ± 0.5 °C.

Samples under high pressure

An optical cell with fused quartz windows was constructed to sustain pressures up to 30 bar. Standard reaction medium was pre-equilibrated with purified oxygen or nitrogen gas under the desired partial pressure for at least 40 min. It was then filled into the optical cell and re-equilibrated with the pressurized gas phase for another 10 min. A small aliquot (typically 50 μ l) of dark-adapted PSII particles from the concentrated stock was injected into the cell (roughly 3 ml liquid, 20 ml gas). The mixture was kept in darkness and equilibrated at the chosen pressure for another 20 min under gentle stirring. After addition, under pressure, of the electron acceptor the sample was excited with a group of laser flashes and ultraviolet absorption transients were recorded.

Laser excitation and spectrophotometry

Samples contained in the pressure-controlled optical cell were excited with a series of flashes from a Q-switched Nd-YAG laser (wavelength 532 nm, duration 6 ns full width at half maximum (FWHM), saturating energy, 100 ms between flashes). Absorption transients were monitored at a wavelength of 360 nm. This wavelength was chosen to reveal charge transfer bands of the Mn_4Ca cluster with minimal contributions from tyrosine, Y_Z (see spectra in Supplementary Information). Data were recorded at an analogue bandwidth of 10 kHz, digitized at 50 μ s per bin, and digitally smoothed when indicated in the figure legends.

Received 26 February; accepted 18 May 2004; doi:10.1038/nature02676.

- Barber, J. Photosystem II: the engine of life. *Q. Rev. Biophys.* **36**, 71–89 (2003).
- Renger, G. Photosynthetic water oxidation to molecular oxygen: apparatus and mechanism. *Biochim. Biophys. Acta* **1503**, 210–228 (2001).
- Robblee, J. H., Cinco, R. M. & Yachandra, V. K. X-ray spectroscopy-based structure of the Mn cluster and mechanism of photosynthetic oxygen evolution. *Biochim. Biophys. Acta* **1503**, 7–23 (2001).
- Zouni, A. et al. Crystal structure of photosystem II from *Synechococcus elongatus* at 3.8 Å resolution. *Nature* **409**, 739–743 (2001).
- Kamiya, N. & Shen, J. R. Crystal structure of oxygen-evolving photosystem II from *Thermosynechococcus vulcanus* at 3.7-angstrom resolution. *Proc. Natl Acad. Sci. USA* **100**, 98–103 (2003).
- Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J. & Iwata, S. Architecture of the photosynthetic oxygen-evolving center. *Science* **303**, 1831–1838 (2004).
- Rutherford, A. W. Photosystem II, the water-splitting enzyme. *Trends Biochem. Sci.* **14**, 227–232 (1989).
- Kok, B., Forbush, B. & McGloin, M. Cooperation of charges in photosynthetic O_2 evolution - I. A linear four-step mechanism. *Photochem. Photobiol.* **11**, 457–475 (1970).
- Hendry, G. & Wydrzynski, T. O-18 isotope exchange measurements reveal that calcium is involved in the binding of one substrate-water molecule to the oxygen-evolving complex in photosystem II. *Biochemistry* **42**, 6209–6217 (2003).
- Hillier, W., Messinger, J. & Wydrzynski, T. Kinetic determination of the fast exchanging substrate water molecule in the S_3 state of photosystem II. *Biochemistry* **37**, 16908–16914 (1998).

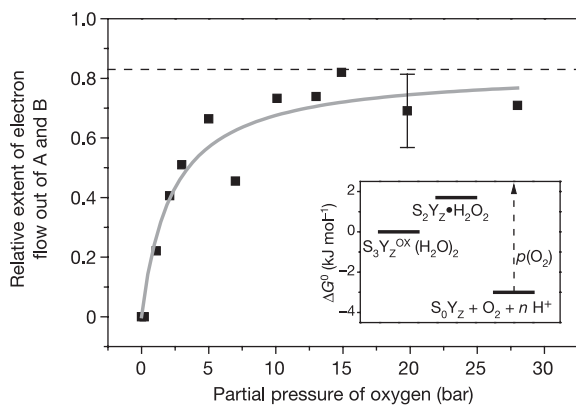


Figure 4 Titration of the corrected absorption transients by elevated oxygen pressure (see text and equation (4)). The insert illustrates the calculated standard free energy profile starting from the highest oxidation state, continuing through the peroxide intermediate, to the fully relaxed state with released oxygen. The error bar denotes the standard deviation of data from 20 experiments with material from six preparations.

11. Renger, G. in *Photosynthetic Oxygen Evolution* (ed. Metzner, H.) 229–248 (Academic, London, 1978).
12. Clausen, J., Debus, R. J. & Junge, W. Time-resolved oxygen production by PSII: Chasing chemical intermediates. *Biochim. Biophys. Acta* **1655**, 184–194 (2004).
13. Babcock, G. T., Blankenship, R. E. & Sauer, K. Reaction kinetics for positive charge accumulation on the water side of chloroplast photosystem II. *FEBS Lett.* **61**, 286–289 (1976).
14. Dekker, J. P., Plijter, J. J., Ouweland, L. & van Gorkom, H. J. Kinetics of manganese redox transitions in the oxygen evolving apparatus of photosynthesis. *Biochim. Biophys. Acta* **767**, 176–179 (1984).
15. Renger, G. & Weiss, W. Studies on the nature of the water oxidizing enzyme. III. Spectral characterization of the intermediary redox states in the water-oxidizing enzyme system Y. *Biochim. Biophys. Acta* **850**, 184–196 (1986).
16. Saygin, Ö. & Witt, H. T. Optical characterization of intermediates in the water-splitting enzyme system of photosynthesis—possible states and configurations of manganese and water. *Biochim. Biophys. Acta* **893**, 452–469 (1987).
17. Haumann, M., Bögershausen, O., Cherepanov, D. A., Ahlbrink, R. & Junge, W. Photosynthetic oxygen evolution: H/D isotope effects and the coupling between electron and proton transfer during the redox reactions at the oxidizing side of photosystem II. *Photosynth. Res.* **51**, 193–208 (1997).
18. Hundelt, M., Hays, A. M., Debus, R. J. & Junge, W. Oxygenic photosystem II: the mutation D1-D61N in *Synechocystis* sp. PCC 6803 retards S-state transitions without affecting electron transfer from Y_Z to P_{680}^+ . *Biochemistry* **37**, 14450–14456 (1998).
19. Karge, M., Irrgang, K. D. & Renger, G. Analysis of the reaction coordinate of photosynthetic water oxidation by kinetic measurements of 355 nm absorption changes at different temperatures in photosystem II preparations suspended in either H_2O or D_2O . *Biochemistry* **36**, 8904–8913 (1997).
20. Rappaport, F., Blanchard-Desce, M. & Lavergne, J. Kinetics of electron transfer and electrochromic change during the redox transition of the photosynthetic oxygen-evolving complex. *Biochim. Biophys. Acta* **1184**, 178–192 (1994).
21. Clausen, J. *et al.* Photosynthetic water oxidation: Mutations of D1-Glu189K, R and Q of *Synechocystis* sp. PCC6803 are without any influence on electron transfer rates at the donor side of photosystem II. *Biochim. Biophys. Acta* **1506**, 224–235 (2001).
22. Lavergne, J. Improved UV-visible spectra of the S-transitions in the photosynthetic oxygen-evolving system. *Biochim. Biophys. Acta* **1060**, 175–188 (1991).
23. Dekker, J. P. in *Manganese Redox Enzymes* (ed. Pecoraro, V. L.) 85–104 (VCH Publishers, New York, 1992).
24. Haumann, M., Drevstedt, W., Hundelt, M. & Junge, W. Photosystem II of green plants: Oxidation and deprotonation of the same component (histidine?) on $S_1^* \rightarrow S_2^*$ in chloride depleted centers as on $S_2 \rightarrow S_3$ in controls. *Biochim. Biophys. Acta* **1273**, 237–250 (1996).
25. Rappaport, F. & Lavergne, J. Coupling of electron and proton transfer in the photosynthetic water oxidase. *Biochim. Biophys. Acta* **1503**, 246–259 (2001).
26. Jahns, P., Lavergne, J., Rappaport, F. & Junge, W. Stoichiometry of proton release during photosynthetic water oxidation: a reinterpretation of the response of Neutral red leads to a non-integer pattern. *Biochim. Biophys. Acta* **1057**, 313–319 (1991).
27. Lavergne, J. & Junge, W. Proton release during the redox cycle of the water oxidase. *Photosynth. Res.* **38**, 279–296 (1993).
28. Junge, W., Haumann, M., Ahlbrink, R., Mulikdjanian, A. & Clausen, J. Electrostatics and proton transfer in photosynthetic water oxidation. *Phil. Trans. R. Soc. Lond. B* **357**, 1407–1417 (2002).
29. Beck, W. F., de Paula, J. C. & Brudvig, G. W. Active and resting state of the O_2 -evolving complex of Photosystem II. *Biochemistry* **24**, 3035–3043 (1985).
30. Klimov, V. V., Allakhverdiev, S. I., Demeter, S. & Krasnovsky, A. A. Photoreduction of pheophytin in the photosystem II of chloroplasts with respect to the redox potential of the medium. *Dokl. Akad. Nauk SSSR* **249**, 227–230 (1979).

Supplementary Information accompanies this paper on www.nature.com/nature.

Acknowledgements We thank H. Kenneweg for technical assistance, H. Heine for advice on the construction of the pressure cell, R. Debus for cooperation on *Synechocystis*, and R. Ahlbrink and A. Mulikdjanian for discussions and help. This work was financially supported by the Deutsche Forschungsgemeinschaft, the Fonds der Chemischen Industrie and the Land Niedersachsen.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.J. (junge@uos.de).

addendum

Measuring fast neutrons in Hiroshima at distances relevant to atomic-bomb survivors

T. Straume, G. Rugel, A. A. Marchetti, W. Rühm, G. Korschinek, J. E. McAninch, K. Carroll, S. Egbert, T. Faestermann, K. Knie, R. Martinelli, A. Wallner, C. Wallner, S. Fujita¹, K. Shizuma², M. Hoshi³ & H. Hasai⁴

¹Department of Statistics, Radiation Effects Research Foundation, 5-2 Hijiya Park, Minami-ku, Hiroshima 732-0815, Japan

²Quantum Energy Applications, Graduate School of Engineering, Hiroshima University, Higashi-Hiroshima 739-8527, Japan

³International Radiation Information Center, Research Institute for Radiation, Biology and Medicine, Hiroshima University, 1-2-3 Kasumi, Minami-ku Hiroshima 734-8553, Japan

⁴Hiroshima Kokusai Gakuin University, 6-20-1 Nakano, Aki-ku, Hiroshima 739-0321, Japan

We wish to clarify misunderstandings created by our Letter on the measurement of Hiroshima fast neutrons¹. Our measurements were partly made within the framework of (and contributed to) a comprehensive reassessment of A-bomb dosimetry conducted by the Joint US–Japan Working Group that produced the new RERF DS02 Dosimetry System, soon to be published^{2,3}. Also, our speculation¹ concerning a “slightly underestimated height-of-burst (HOB) for the Hiroshima bomb” does not imply that the HOB should be changed based solely on the ⁶³Ni measurements. Although our work provides direct information on dose-relevant fast-neutron fluence, it should not be construed to be the sole basis for resolution of the Hiroshima neutron discrepancy that had been reported for thermal neutrons. The original authors were not fully aware of the scientific input of the copper sampling in Hiroshima and wish to remedy this here by extending the author list and by acknowledging additional support from Japanese funding agencies (Grants-in-Aid for Scientific Research B and C of the Japan Ministry of Education, Culture, Sports, Science and Technology for support and the Radiation Effects Research Foundation (RERF) in Hiroshima for the copper sampling effort). □

1. Straume, T. *et al.* Measuring fast neutrons in Hiroshima at distances relevant to atomic-bomb survivors. *Nature* **424**, 539–542 (2003).
2. Cullings, H. M. & Fujita, S. The way to DS02: resolving the neutron discrepancy. *RERF Update* **14**, 17–23 (2003).
3. DS02: A New Dosimetry System for A-Bomb Survivor Studies Joint Senior Review Group Report on DS02 (<http://tis.esh.doe.gov/health/ihp/japan/DS02.pdf>) (March 2003).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel +49 89 54 90 57 11/-2

Fax +49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: r.asami@naturejpn.com

naturejobs

Alternative paths

The two recipients of this year's entrepreneur awards, given out last week by the London Biotechnology Network (LBN), provide good examples of different routes into business. One pursued a commercial career from the word go, undertaking directed training and choosing jobs in various commercial fields. The other began as an academic but found that he had to spin out a company in order to move his research closer to the clinic.

The winner of the LBN's Entrepreneur of the Year Award, Sanjay Kakkar, chief executive of London biotech firm Trigen, knew early on that he wanted to go into industry. After completing his medical degree, he earned a master's in healthcare management at Harvard University and began his commercial career by joining the venture-capital subsidiary of the Bank of Boston in London. He then moved on to stints with Sandoz (now Novartis) in Milan, Italy, and Pharmacia in the United States.

By contrast, Robert Coffin, the recipient of the Young Entrepreneur of the Year Award, launched his vaccine company BioVex from academia. The company emerged out of research done at University College London's Institute of Child Health. Coffin still maintains a balance between both worlds — he lectures at the university and serves as chief scientific officer at the Oxford-based company.

The LBN awards are not just beneficial for their recipients. They show young scientists in Britain and beyond that viable careers are possible in biotechnology, no matter what your age and career path. And they illustrate the vitality of the biotech sector in London, where about 70 of the area's 100 or so companies are less than five years old. Recognizing pioneers in these companies may prove to be a great way to get scientists involved in British biotech — whether they come from an industrial or an academic background.

Paul Smaglik
Naturejobs editor



Contents

SPECIAL REPORT

Making the move into
drug sales p486

CAREER VIEW

Recruiters & Industry
Dual competencies
Graduate Journal
Moving on
Movers
Mark Willenbring p488

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Ever more scientists are joining drug companies' sales departments. Are they happy there?

Marika Willerroider investigates.

The young man in the doctor's waiting room checks his watch. He is not a patient, but he has spent more than an hour here now, and he has five more physicians to visit this afternoon. Inside the surgery, the doctor is also feeling a bit hassled. It is his fifth appointment with a medical representative that day.

Physicians and the pharmaceutical industry's travelling salespeople have a partnership of convenience. The drug market is competitive, and the widespread practice of tempting clients with presents, or even free conference trips, carries a whiff of corruption. But a job as a medical representative — a fancy name for a door-to-door drug seller — is often the first step in a lucrative career in the sales or marketing departments of big drug companies, and hence is an appealing alternative for an increasing number of scientists tired of the insecurity in research.

Gerd Sauer is one such case. The 41-year-old molecular biologist has a PhD from the University of Würzburg in Germany. Ten years ago, seeing no future at the bench and feeling his appetite for research waning, he swapped his lab coat and pipettes for a suit, briefcase and company car, and became a sales rep at Immuno, a drug company that later merged with Baxter Healthcare.

To researchers this might seem a step down. But Sauer, now group product manager for oncology at Roche Pharma in Basel, Switzerland, doesn't regret his decision. "Leaving research was not easy, but it proved to be the right step for me," he says. "I wouldn't like to have missed one second of the time I have spent working in sales. In the lab you have to be fortunate to get good results; in sales you are responsible for achieving your goals." And meeting interesting people and being encouraged to produce new ideas provided the motivation to turn his back on science.

In Germany, some 17,000 medical representatives look after about 130,000 physicians. In France, Britain and the United States, there are similarly about seven doctors to every salesperson. And many of the reps



come from science. At big companies such as Roche or GlaxoSmithKline, up to 40% of sales and marketing staff have a PhD, most in biology, pharmacology or medicine.

"The job market for researchers forces many graduates and life scientists into the pharmaceutical industry," says Peter Eulenschläger, head of human resources at GlaxoSmithKline's German headquarters in Munich. Not all former scientists find their dream job, but those with good communication skills, who like contact with people and being out on business, have good chances to progress.

A QUALIFIED OPINION

Management careers in the drug industry typically begin outside the office. Starting salaries for sales reps are roughly equivalent to those for postdocs, but bonus payments for hitting sales targets push them much higher.

A PhD is not essential, but it is appreciated by most recruiters, and can be helpful in a medical representative's daily routine. After all, most physicians prefer to be looked after and advised by well-trained specialists. "A pharmacologist or life scientist will simply be able to explain much more professionally and competently the pros and cons of the products he



On the road: newly trained medical representatives can expect to spend their first two years working out of the office, selling drugs to doctors.

Aventis in Frankfurt, Germany. "To be successful in this business you need at least as much determination, intuition and enthusiasm as in science."

"It makes no sense trying this if you know from the start that you don't like business," adds Sauer. "A salesperson has to be authentic."

After two years or so outside the office, successful medical representatives can expect a promotion to a sought-after management position, such as product sales manager or marketing group leader, says Thomas Hefti, a marketing director at Roche in Basel.

But scientists often regret their decision early on. "I thought to myself: this is not the kind of work that I would like to do for the rest of my life," says Matthias Strobl, a PhD student at the Institute of Nephrology in Munich, who after graduating took up a job in the marketing department of what was then SmithKline Beecham Germany, but decided to quit and return to the lab.

"I simply wasn't made for this," says Martin Klein, a biologist by training, who quit

his job as a sales representative after a few months. "I felt totally overqualified, but then again patronized by the physicians I dealt with." Frustrated, Klein became a social worker for adolescents in Salzburg, Austria.

Sometimes, says Klein, he felt like "Santa Claus in a suit", as he piled gifts on customers' desks. Meanwhile, however, the drug industry is trying to get rid of such practices. "In the past few years, pharmaceutical companies have been trying hard to improve their image and customer relations," says Roland Stahl, a spokesman for the German Association of Health Insurance Doctors. "The emphasis has really shifted from pushy sales practices towards serious information of, and advice to, doctors."

In most European countries this shift is backed up by laws and guidelines that regulate the number and value of gifts and other benefits that physicians can accept from drug companies. Sales managers such as Sauer, who are keen to develop a science-led, service-orientated relationship with their customers, support this trend, which they feel enhances the value of their profession. It would seem that the days are over when a physician could reply to the offer of a free conference trip with: "Oh, I would love to go — if it's in the Caribbean and I can bring my partner."

■
Marika Willeroeder was an intern in *Nature's* Munich Office; she recently finished a PhD in genetics at the University of Salzburg, Austria.

or she is trying to sell," says Axel Munte, a Munich-based physician and former head of the Bavarian Association of Health Insurance Doctors.

Indeed, many physicians feel swamped by new drugs and clinical studies. Because few have time to scan even the most essential literature, competent medical reps can keep practitioners up to date.

"The pharmaceutical industry is an essential source of further education for physicians," says Joe Collier, the London-based editor of the *Drug and Therapeutics Bulletin*, a review published by the UK Consumers' Association. "It seems appealing therefore that drug-company representatives should be well-trained scientists who know what they're talking about."

Still, a sales rep's primary task is to sell products and make money. And for some former scientists, who may have gone for days in the lab without speaking to other people, and who normally have little experience with corporate culture and business practices, this can be rather difficult.

When they join the sales and marketing departments of companies such as Roche, newcomers usually get three to six months of training before they start visiting customers.

"This is really a new challenge," says Claus-Henning Albrecht, who, after having spent several years in medical research, last year began a traineeship at

GRADUATE JOURNAL

Moving on

With a sense of sadness, I write my last journal entry from the United States. I'm moving on. In a month I'm going back to London where my journey began.

When I departed for the United States, I left behind a relationship, family and friends, a familiar life and a country I knew. Boarding that plane, I was both apprehensive and excited. I had planned to stay for a year. In the end, I stayed for two.

Needless to say, making a new life in another country isn't the easiest of tasks. But I very quickly settled in, made a home for myself and found new friends.

The experience has taught me many things, and is one I won't forget. What will I miss the most? Without doubt, the people. My wild and wacky physicist friends will be with me for ever! Their kindness, generosity and *joie de vivre* have taught me to be a better person. It is a rare treat to work with people from all over the globe and to discover friendships that cross the boundaries of culture. Four o'clock teatime will never again be the same without at least six different nationalities round the table...

Leaving your roots isn't easy. Neither is going back. But there are many things to tempt me home: good food, the BBC, public transport... And so a new journey begins. See you on the other side of the Atlantic. ■

Amber Jenkins is a graduate student in particle physics at Imperial College, London, doing thesis research at Fermilab in Batavia, Illinois.

RECRUITERS & INDUSTRY

Dual competencies

As biotechnology companies grow, they tend to develop their operational and management team by bringing in people who have dual competencies — formal qualifications in both science and business.

Such skill sets are attractive to companies because they allow them to expand their existing competencies without needing to employ a huge number of people. They can also aid communication between research-based innovation and administrative support.

For scientists, having dual competencies can accelerate their career path — especially in the private sector. This tends to mean greater management responsibility — and usually higher salaries — than their bench-bound colleagues.

But for many of these high-fliers, the move into

management will spell the end of lab work, leaving them facing an office-based future. Indeed, researchers and engineers who will have studied for 5–8 years for their scientific qualifications very often quit their lab-based posts after just one year of business school. This is a sad situation, as the skills that they are gaining in business readily translate into the world of research.

For example, a knowledge of accounting and finance is essential in product development for understanding investment, value and the return on investment. Similarly, skills in the management of human resources will aid the planning and recruitment processes to get projects up and running.

At Protein'expert, a biotech firm in Grenoble, France, we have sought to tackle this with our strategy of employing or shaping (on-the-job education)

people with dual skill sets. Of the six members of the management team, five have both science and business qualifications — the exception is the marketing manager. Although both the chief executive and director for intellectual property and quality control have left the bench behind them, the other three dually skilled managers all make use of their newly acquired business skills to guide scientific projects.

Whether they are used to change career direction, or simply to enrich the current path, dual competencies represent a challenging opportunity for scientists. The acquisition of business skills should be encouraged — both at universities and in the workplace as they can bring huge value to scientific projects. ■

Tristan Rousselle is chief executive of Protein'expert in Grenoble, France.

♦ www.proteinexpert.com

MOVERS

Mark Willenbring, director, division of treatment and recovery research, US National Institute on Drug Abuse and Alcoholism, Bethesda, Maryland



Despite Mark Willenbring's best efforts early in his career to avoid becoming a specialist, the psychiatrist soon found himself identified as the 'alcohol guy'.

During his training, internship and residencies, Willenbring had liked the idea of exploring his broad interests in clinical psychiatry as an academic, splitting his time between teaching, administration and research. But in the last year of Willenbring's residency, while he was seeking his next post, a

colleague suggested that he should consider moving into the treatment and research of alcohol and drug addiction. He was told that the National Institutes of Health (NIH) was offering career teaching grants in alcohol and drug-abuse research. Willenbring applied and, although in the end he decided not to accept the grant, going through the process got him pegged as an "alcohol and drug expert", he says.

This reputation helped him to secure a fellowship at the University of Wisconsin in Madison, where he got involved in developing a curriculum in addictions and their treatment. Doing that helped him to become the specialist that the university thought he was and that he had initially resisted becoming.

By then up to speed in addiction psychiatry, Willenbring further expanded his experience by working at the Dane County alcohol detoxification centre. This reshaped his view of alcoholism,

leading him to see it as a chronic illness. As a result, he sought ways to treat the condition rather than trying to find a complete cure. He describes his major research focus as "populations of people everyone has given up on".

Willenbring, who last month moved from the University of Minnesota to become director of treatment and recovery research at the US National Institute on Alcohol Abuse and Alcoholism, says that when he began his career, the market trend towards hiring specialists was just beginning. He wonders now whether broad-based experience should be viewed in a more positive light — especially as science and medicine are becoming more interdisciplinary.

A mixture of experiences can be helpful if you have to manage large projects. And Willenbring will be drawing on his generalist beginnings as he finds his feet as the 'alcohol guy' at the NIH. ■

CV **1982–2004:** Veteran's Affairs Medical Center, Minneapolis, Minnesota, rising to director, addictive disorders section (1991)
1980–82: Medical director, Dane County Alcohol Detoxification Unit, Madison, Wisconsin
1980–82: Staff psychiatrist, Dane County Mental Health Center, Madison, Wisconsin
1977–80: Staff physician, Kaiser Medical Center, Sacramento, California